教师基本功丛书·数学教师卷

编著 许世红 胡中锋

数学试卷
分析方法

ShuXueShiJuan FenXiFangFa



华东师范大学出版社





图书在版编目(CIP)数据

数学试卷分析方法/许世红,胡中锋编著. —上海:华东师范大学出版社,2009
(教师基本功丛书·数学教师卷)
ISBN 978-7-5617-7210-2

Ⅰ.数… Ⅱ.①许…②胡… Ⅲ.数学课—教学法—中小学 Ⅳ.G633.603

中国版本图书馆 CIP 数据核字(2009)第 177005 号

教师基本功丛书·数学教师卷
数学试卷分析方法

编　　著　许世红　胡中锋
策划组稿　李文革
审读编辑　李文革
封面设计　黄惠敏

出版发行　华东师范大学出版社
社　　址　上海市中山北路 3663 号　邮编 200062
电话总机　021-62450163 转各部门　行政传真 021-62572105
客服电话　021-62865537(兼传真)
门市(邮购)电话　021-62869887
门市地址　上海市中山北路 3663 号华东师范大学校内先锋路口
网　　址　www.ecnupress.com.cn

印 刷 者　华东师范大学印刷厂
开　　本　890×1240　32 开
印　　张　7.75
字　　数　202 千字
版　　次　2009 年 10 月第一版
印　　次　2009 年 10 月第一次
印　　数　3100
书　　号　ISBN 978-7-5617-7210-2/G·4168
定　　价　15.00 元

出 版 人　朱杰人

(如发现本版图书有印订质量问题,请寄回本社客服中心调换或电话 021-62865537 联系)

教师基本功丛书
数学教师卷

数学
试卷分析方法

华东师范大学出版社

前　言

本书的读者群为中小学教师。

试卷分析包括“试卷质量分析”和“测验成绩分析”两大部分。进行试卷分析所依据的测量理论不同，所得的结论也会有较大的差异。

目前应用广泛的测量理论主要有：(1)经典真分数理论，(2)概化理论，(3)试题反应理论。每种理论都有其应用范围、各自的优点与不足。考虑到中小学教师主要了解经典真分数理论，而且平时使用的试卷分析也主要基于经典真分数理论，本书主要介绍经典真分数理论及其应用。概化理论和试题反应理论则只做简要介绍，目的是希望一线教师对测量理论的发展有所了解。

本书共计五章。第一章是本书的理论基础，包括三节，主要介绍经典真分数理论的基本内容、优点与局限，以及应用时应注意的问题；教育测验的基本常识；教育统计的基本常识。概化理论和试题反应理论的介绍则放在附录部分，供查阅。第二章介绍试卷分析中的命题质量分析，结合具体案例详细剖析了四个基本技术指标(信度、效度、难度、区分度)。第三、四章侧重于介绍试卷分析中的测验成绩分析，其中第三章主要介绍运用描述统计方法(统计图、统计表、集中量数、差异量数等)对测验成绩进行组织、表达、整理与概括；第四章主要介绍运用常见的基本推断统计方法(相关分析、回归分析、方差分析)对经过整理与概括的测验成绩做较为深入的分析，为后继教与学的决策提供依据。第五章依据前四章介绍的基本测量理论、统计和评价技术，以案例分析的方式介绍中小学数学教学

中常见测验的试卷分析方法，重点放在根据教师自编测验卷或统考测验卷，运用测验分析技术，协助教师掌握基本的数据收集、处理、分析方法，用于诊断教师教学、学生学习状况，以便改进以后的教学，并尝试给出实用的试卷分析报告模板。

目前中小学教师接触较多的数据处理软件主要有 EXCEL 软件与 SPSS 软件，两种软件操作都比较方便。为了方便教师更好地学习与使用书中的技术与方法，本书详细呈现使用 EXCEL 软件与 SPSS 软件计算命题质量技术指标、制作统计图表、计算统计量、进行基本统计分析的具体操作步骤，努力体现实操性、应用性，力争为中小学教师提供工作便利。

考虑到本书的读者是中小学教师，主要从事实际的教学工作，难以有时间从头到尾集中阅读，使用本书时建议读者从书中案例入手，在模仿与操作的过程中理解基本的测量与评价专业知识，然后再学习并加深对有关理论的认识。

本书的结构框架由许世红、胡中锋共同构思，许世红完成初稿，胡中锋审阅全部稿件。广州大学附属中学施永红老师、广州市五羊中学高惠平老师、广州市文德路小学黄丽芳副校长等为本书案例提供了详尽的素材与数据，并审阅了有关章节，在此谨致谢意。

本书编著过程中，参阅了不少文献，并引用了其中一些资料，有些未在正文中一一标注，谨在书后参考文献中列出，一并表示感谢。

由于时间仓促，书中难免存在瑕疵，祈请各方专家与同行不吝赐教，并欢迎广大读者批评指正。

许世红　胡中锋

2009 年 8 月于广州

目　　录

第一章　绪论 1

第一节　经典真分数理论 1

一、能力及其可测性 1

二、测验分数的真分数模型 2

三、真分数理论的优点 5

四、真分数理论的局限 6

五、应用真分数理论应注意的问题 8

第二节　教育测验简介 9

一、教育测验的基本特征 9

二、教育测验的功能 11

三、常模参照测验与标准参照测验 13

第三节　教育统计简介 16

一、教育统计的含义 16

二、教育统计的内容 16

三、教育测验与教育统计的关系 17

第二章　试卷质量的基本分析 19

第一节　测验信度 20

一、信度概念 20

二、测验信度的估算方法 21

三、影响测验信度的因素 24

四、测验信度的定性评价指标 25

五、测验信度的定量评价标准 28

第二节　测验效度 29

一、效度概念 29

二、效度与信度的关系 31

三、效度分类与评估…………………………………………………………32
四、测验效度的定性评价指标………………………………………………36
第三节　测验难度………………………………………………………………37
一、难度概念…………………………………………………………………38
二、难度的计算………………………………………………………………38
三、难度的评价标准…………………………………………………………39
四、影响难度的因素…………………………………………………………40
五、试题难度的定性评价……………………………………………………41
第四节　测验区分度……………………………………………………………41
一、区分度的概念……………………………………………………………41
二、区分度的计算……………………………………………………………42
三、区分度的评价标准………………………………………………………44
四、区分度与难度、信度的关系……………………………………………45
五、影响试卷区分度的因素…………………………………………………47
第五节　EXCEL 与 SPSS 软件应用实例……………………………………47
一、计算测验信度……………………………………………………………47
二、计算测验效度……………………………………………………………53
三、计算试题的难度…………………………………………………………62
四、计算试题的区分度………………………………………………………64

第三章　测验成绩的统计处理……………………………………………79
第一节　数据的特点与种类……………………………………………………79
一、数据的常见三种分类……………………………………………………79
二、表示测验成绩的数据的基本特点………………………………………81
第二节　测验分数的组织与表达………………………………………………82
一、茎叶图……………………………………………………………………83
二、频数分布表………………………………………………………………85
三、频数分布直方图…………………………………………………………87
第三节　测验分数的图表表示…………………………………………………88

一、图形的特点……88
二、常用统计图……89
三、统计表制作的规则……93
四、常用统计表……95
第四节 测验分数的概括……96
一、集中量数……96
二、差异量数……98
三、分数分布的形状……100
第五节 EXCEL 与 SPSS 软件应用实例……102
一、对测验分数进行初步整理……102
二、计算描述性统计量……109
三、制作统计图表……115

第四章 测验成绩的统计分析……126
第一节 相关分析……126
一、相关与相关系数……127
二、相关系数的类型及其计算……129
三、相关系数的应用……136
第二节 回归分析……137
一、回归分析与相关分析……137
二、回归分析的主要步骤与基本类型……138
三、一元线性回归……140
四、多元线性回归……149
第三节 方差分析……151
一、方差分析的基本原理……152
二、单因素方差分析……154
三、双因素方差分析……162
第四节 EXCEL 与 SPSS 软件应用实例……164
一、相关系数计算与显著性检验……164

二、回归方程计算与有效性检验……170
三、方差分析与平均数差异检验……178

第五章　试卷分析报告的基本模式……191
第一节　试卷分析报告的基本框架……191
一、整卷层面……191
二、题组层面……195
三、试题层面……198
第二节　三种常见测验的试卷分析报告基本模式……201
一、单元测验……201
二、学期测验……207
三、联考或统考……211

附录 1　概化理论简介……215
附录 2　试题反应理论简介……224

主要参考文献……234

第一章
绪　论

现代教育发展至21世纪，检验基础教育阶段一线教师教学基本功底情况，虽然还是从“备课、上课、批改作业、出测验卷、进行试卷质量分析”等环节着手，但各个环节的评价内容、评价标准已经有了巨大的变化。教育测量在检验教学效果、评估学业成就等方面应用极为普遍，掌握基本的测量理论并正确应用测量理论检验、反馈、指导教学活动的开展成为现代合格教师应该具备的基本技能。

本章主要从真分数模型、真分数理论的优点、真分数理论的缺陷、真分数理论应用时应注意的问题等方面介绍在日常教学实践中应用广泛的经典真分数理论。另外，结合学校实际需要，简要介绍教育测验、教育统计的一些基本常识。

第一节　经典真分数理论

一、能力及其可测性

在日常生活中，人的身高、体重等特征都比较容易测量，因为这些生理属性比较稳定、直观，所需要用的测量工具也容易制作和使用。人们也尝试去测量诸如天赋、智力、动机、性格等方面的人的心理特征，在教育测试中，我们常常把要测量的这种心理特征称为能力。能力即完成一组特定认知任务所表现出的相对稳定的思维或

行为方式，由于这种能力往往与学生学习的具体科目相联系，因此，就有了数学能力、语文能力、物理能力等之分，有时候也会把某种能力进行分解，例如数学能力进一步分解为计算能力、推理能力、空间想象能力等。

通常，我们假定人与人之间的能力有差异，这种差异与人们完成一定难度任务的正确程度相关。一般情况下，对有一定难度的任务，能力高的人正确完成的概率高，能力低的人正确完成的概率低。类似地，对于具有一定能力的人，正确完成高难度任务的概率较低，而正确完成低难度任务的概率较高。

由于能力没有明确的物理与生理属性，无法直接测量，因此人们设计出特定的测量量表（如数学测验卷），通过分析考生解答测验卷题目的过程与结果情况（如考生数学测验的答卷情况）来推测考生具备某种能力（如某个阶段某类数学能力）的特点与水平。

需要明确的是，无论教育考试测量的是何种能力，考试测量得到的数据只是考生某个方面能力的一个估计值，单凭这个估计值往往无法全面衡量考生的真实学习状况。

二、测验分数的真分数模型

为了了解学生的学习能力状况，我们组织考试对学生进行考查。经过考试测量后，学生的学习能力体现为一个数值。然而，由于测量误差的存在，实际测得的数值往往与学习能力的真实水平不完全一致。例如，我们常常说“××考生基本上考出了应有水平”、“××考生这次测验超水平发挥”、“××考生的学习状况基本上测出来了”等，就是对这种测量现象的一种描述。

为了研究方便，我们把反映学生学习能力真实水平的那个数值称为学习能力的真分数，把实际测量得到的分数称为学习能力的实测分数，把实测分数与真分数之间的差值称为测量误差。当实测分数与真分数很接近时，就说这次测量的误差很小。由此，我们得到关于真分数假设的数学模型：

假设 1　$x_i = t_i + e_i$。　(1.1)

其中，x_i 表示在某次考试中考生 i 能力的实测分数，t_i 表示考生 i 能力的真分数，e_i 表示考试中的误差分数。

真分数的数学模型 1.1 与人们在物理量测量中的感性经验相吻合，而且表达式为和式，在数字计算中使用非常方便。

与模型 1.1 同时提出的还有 3 个相关联的基本假设：

假设 2　真分数不变，是一个常数。

在讨论具体的能力测试时，我们假定每个考生的真实能力是不变的，是一个确定值。在研究一大群考生时，一般情况下，假定考生的真实能力的分布状况服从正态分布。

假设 3　反复施测时，误差分数相互独立，且服从期望为零的正态分布。

考生的能力真分数是无法直接得到的。根据假设 3，同一测验反复施测时，误差分数呈零相关，即 $\rho(e_1, e_2) = 0$。既然误差分数取值是服从期望为零的正态分布（即 $E(e) = 0$），那么，如果能够求出误差分数方差，就能以实测分数为中点，以误差分数标准差为半长，对能力真分数取值进行区间估计。

假设 4　用平行测验对同一总体考生施测后，所得实测分数的平均数与方差相等。

从理论上讲，如果一组测验测量的是同一种能力，那么在控制方差能力相同的情况下，这组测验所测得的实测分数就应具有相同的分布，即不同测验的实测分数具有相同的平均数与标准差，这样的一组测验就称为平行测验。

根据上述四个假设，我们立即可以得出以下几个重要结论，这些结论构成了真分数理论的基石。

推论 1　真分数等于实测分数的期望值，即 $t = E(x)$。

推论 2　真分数与误差分数彼此独立，即 $\rho(t, e) = 0$。

推论 3　实测分数的方差等于真分数方差与误差分数方差之和，即 $\sigma_x^2 = \sigma_t^2 + \sigma_e^2$。

推论4 平行测验的真分数平均数相等，真分数标准差相等。

推论5 平行测验向考生总体施测后，个体内的误差分数标准差等于每个测验的误差分数标准差。

推论6 在一组平行测验中，任意两个平行测验间的相关均相等。

【例1-1】 根据上述假设与推论，我们构建一个施测的模型。假设有一个考生总体，它由 m 个考生构成，这 m 个考生参加了 n 次平行测验，得到的实测分数构成表1-1所示的数字矩阵。

在表1-1中，考生 i 参加一组平行测验所得的实测分数 x_{ij}（$i=1, 2, \cdots, m$）可以表示成 $x_{ij}=T_i+E_{ij}$，$j=1, 2, \cdots, n$。其中，考生 i 参加一组平行测验的个体真分数是一个确定值 T_i，它不随测验的改变而改变。由于这组平行测验测量的是同一种能力，因此每个测验的真分数平均数都是 $\overline{T}$，真分数标准差都是 σ_T。因为误差方差大小仅取决于测验控制误差的能力，且平行测验假定误差控制能力相等，所以，单一测验内的个体间误差分数方差等于这 n 个平行测验向某个考生施测后所得误差分数方差，即标准差都相等，均是 σ_E。

表1-1 一群考生参加一组平行测验所得分数矩阵

实测分数		n 个平行测验						个体真分数	误差分数	误差分数标准差
		1	2	…	j	…	n			
m 个考生	1	x_{11}	x_{12}	…	x_{1j}	…	x_{1n}	T_1	$\{E_{1j}\}$	σ_E
	2	x_{21}	x_{22}	…	x_{2j}	…	x_{2n}	T_2	$\{E_{2j}\}$	σ_E
	⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮
	i	x_{i1}	x_{i2}	…	x_{ij}	…	x_{in}	T_i	$\{E_{ij}\}$	σ_E
	⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮
	m	x_{m1}	x_{m2}	…	x_{mj}	…	x_{mn}	T_m	$\{E_{mj}\}$	σ_E
测验真分数	平均数	$\overline{T}$	$\overline{T}$	…	$\overline{T}$	…	$\overline{T}$			
	标准差	σ_T	σ_T	…	σ_T	…	σ_T			
测验误差分数标准差		σ_E	σ_E	…	σ_E	…	σ_E			

根据表1-1，如果测验对误差控制得很好，那么用这n个平行测验向一组考生施测，所得的实测分数就会毫无偏差地传达考生真实能力的信息，实测分数实际上就是表达在另一量尺上的真分数的线性变换值，这时，就称测验性能良好。因此，我们可以通过研究两个平行测验所得的实测分数之间的相关性，分析测验结果的一致性与可靠性，以及测验误差的控制情况，这些相关性研究就是所要估计的各种可能信度指数、信度系数等问题，因此，人们也称真分数理论为经典信度理论。

除了信度估计外，真分数理论也探讨其他相关的测量学问题，如：效度、测验编制、常模、测验等值、测验偏差、试题分析、常模参照测验、标准参照测验、适应性测验、题库建设等等，本书主要应用真分数理论研究试卷的分析方法，涉及的主要是真分数理论中有关信度、效度、试题分析等部分的基本概念与基本方法。

三、真分数理论的优点

经典真分数理论属于早期心理计量学理论，它形成于20世纪初，成熟于20世纪50年代，其基本目的是要形成一种可操作的估计测验分数测量精度的方法，它的数学模型简单易懂，在实际工作中应用广泛，在我国的教育测量与评价实践中仍占据主要地位。具体而言，该理论的优点主要体现在以下三个方面。

1. 以弱假设为基础，突出主要矛盾

真分数理论产生之初，模仿物理测量的思路，按照“真值就是反复测量的期望值”构建计量模型，突出“主要通过外部控制测验误差、提高实测分数的精度，从而提高对真实分数的估计精度”的研究主线，抓住主要矛盾，弱化其他的问题（如对误差来源的具体分析，对测验分数的单位等值要求，对平行测验的严格检验等），对学业成就进行整体性笼统的总分评价，通过将考生简单地当做一个整体在同一量尺上排队加以分析，在测验编制、分数解释、效度验证等具体工作中发挥了重要指导作用，较好地满足了实际工作的需要。

2. 采取概率的观点构建模型，把心理变量间的关系视为随机变量间的关系

真分数模型 $X = T + E$ 中，实测分数 $X = (x_1, x_2, \cdots, x_n)$，真分数 $T = (t_1, t_2, \cdots, t_n)$ 与误差分数 $E = (e_1, e_2, \cdots, e_n)$ 都是随机变量。平行测验中，由于随机因素的影响，测验结果往往会有不同的数值，采用随机变量就可以描述平行测验中测验结果的差异性，这样，既可以借助概率知识研究测验结果的规律性，又可以借用数学函数等工具刻画测验中各因素之间的内在联系。这种研究方法反映出心理与教育测量的科学取向。

3. 进一步拓展了对测量效度、测验公平性与测验等值等问题的探索

随着社会、政治、经济、文化的发展，人们对测验内容、测验功能、测验适用面、测验精度等的要求越来越高，真分数理论也根据实际需要自觉地开展了相关问题的理论与实践研究，贡献出特定而且有力的概念、原理与方法，因此，真分数理论仍显示出很强的生命力。

四、真分数理论的局限

虽然真分数理论无论在理论的基础研究方面还是在实践的具体指导方面，均为心理与教育测量的发展做出了巨大贡献，但是它的理论框架存在明显的先天缺陷，在测验实践飞速发展的今天已经日益显示出其局限性。

1. 计量模型建立在考生答题的外在表现上，导致各项测验性能指标严重依赖样本

真分数模型 $X = T + E$ 中，实测分数 X 反映的是考生在测验中的外在具体表现，在具体分析误差分数 E 时也侧重于研究由外在因素引起的随机误差的控制，因此，基于真分数计量模型而界定的测验信度、效度、区分度、难度等测验性能指标也依赖于考生群体的具体表现。例如，真分数理论把测验信度定义为实测分数与真分数的

相关性，而相关系数的计算受样本数据取值全距范围大小的影响，因此也就极其依赖样本；又如，真分数理论把试题难度定义为考生样本组上的通过率，如果考生样本组的水平高，则难度就小，反之，则难度大，难度的计算结果严重依赖样本组。这样，测验的结果只能推广到与考生样本组非常类似的群体中，其适用推广范围非常有限。

2. 所测能力的不变性、反复测试中误差分数相互独立性、严格平行测验等假设在实践中难以满足

真分数理论源于物理测量的研究，但是心理与教育测量的对象是人，它与物理测量对象"物"之间有着本质的区别。首先，所测的对象都是内隐于个体内部的特质，测量所能直接观测的只是这些内部特质的外在表现，因此，只能是间接测量。其次，测量的过程中，考生的记忆、遗忘、发展新技巧等心理因素都会主动产生作用，所以，多次施测甚至前一部分试题的施测，都会对考生产生启发与练习等作用，进而对后续测验与试题产生影响。另外，测试题目的代表性、测验分数的评定、施测的时空条件等因素，也会带来影响。这样，真分数理论的基本假设就很难满足，基于真分数理论的数据分析就显得很粗糙，误差较大。

3. 基于严格平行测验等定义的测验信度及相关计算较为粗略

真分数理论中，测验信度的界定建立在平行测验的基础上。由于实际操作中平行测验难以满足，因此测验信度也就不够准确。另外，利用同一测验采用重测、复本、折半等不同方法计算可以得到多个信度系数，结果就出现了多个测量标准误值的尴尬局面。

4. 试题难度与考生能力水平值定义在不同的量尺上

真分数理论的试题难度是考生组的通过率，参照系是某一个考生样本组。而考生能力水平值用测验的总得分（或试卷中的试题通过率）来表示，参照系是某一组试题。由于考生样本组与某一试题组彼此互不隶属，因此，一道试题的难度为0.55，只能解释为某一个考生样本组中有55%的考生答对了该试题，却不能反映答对该题目

的考生能力的高与低，即试题难度与考生能力建立在不同的量尺上。事实上，人们希望试题的难度应反映考生正确作答的能力水平，通过试题难度可以直接比较考生的能力高低，而真分数理论显然没有解决这一问题。

五、应用真分数理论应注意的问题

虽然真分数理论存在诸多的局限性，但是由于该理论通俗易懂，可操作性强，在我国的教育实践领域应用很广泛。应用真分数理论开展教育测验与评价时，需要注意以下几点问题。

1. 测验卷制作应体现较为清晰的参照标准

每一种教育测验都有其特定的目的、功能及适用范围。例如，常模参照测验以全体考生在某一大规模测验中所得到的成绩分布为标准，衡量考生在这一测验上所得到的成绩在全体成绩分布中所处的地位，它主要适用于大规模的选拔性考试，如中考、高考等；标准参照测验是以考生对测验目标或内容的掌握程度作为标准，对测验分数进行解释，如学科的单元测验、学期测验等。在编制测验卷时，要明确测验的性质与功能，确保测验使用与解释的合理性。

2. 实施测验时尽量避免无关因素干扰

为了保证测验结果的有效性与可靠性，除了把好测验卷编制质量关外，严格控制测验实施过程，确保实施过程的公平性与合理性，也有利于考生在完成相应的测验时顺利展示应有的能力与水平。因此，实施测验时，应按照测验实施规定的要求进行，严格控制这一过程中可能出现的各种无关因素的干扰，如测验卷的发放、考试时间的控制、作弊行为的监控、测验环境的选择等，确保测验结果的可靠性。

3. 正确认识测验结果

人们总是希望测验结果是考生能力的最好测度，在解释测验结果时，往往将测验分数的高低与考生能力的高低直接对等，仿佛这种做法，即根据测验结果评价考生，比其他评价方法显得更加公平、

公正。但从教育测量角度来看，对测验结果的认识需要综合考虑多种因素。首先，测验结果除了与考生在学校学习状况相关外，还受家庭情况、考生生理与健康情况等因素的影响，如果测验结果出现异常，需要从多个角度寻找原因。其次，测验结果的解释应放在考生群体中进行分析，脱离考生群体单独谈论某一个测验分数的高低，其价值很有限。另外，测验结果是对考生能力的间接反映，它受限于试题命制状况，也受评卷过程影响，解释测验结果时，也需要交待这些相关因素。

第二节　教育测验简介

由于使用目的与需求的不同，测验发展至今，已经有多种分类。本书主要探讨学校教学领域中的试卷分析方法，相应地，探讨的重心限于教育测验。本书中的教育测验指学生在学校接受学科学习与训练后，对学生所获得知识、技能与能力状况进行考核，并将考核结果转化为数量描述的过程。

一、教育测验的基本特征

教育测验的对象是学生掌握相关知识、技能与能力的状况，研究的是学生的高级心理活动，这种心理活动是内隐的，它不能借助高精密仪器直接认识，因此，它与人们经常接触的物理测量不同。教育测验主要具有以下 4 个基本特征。

1. 测验方法与结果表示的间接性

教育测验的间接性体现在两个方面。一是测验方法的间接性。人们无法直接测量考生的内在心理特性，而只能通过其外显的行为，来间接测量其心理活动的特征与水平。例如，如果想考查考生某方面的数学能力，只能通过考生对某类测验题目的反应和一些行为表现，借助间接获取的结果，运用推理、判断的方法，去大致推断

考生知识掌握水平与能力高低。二是测验结果表示的间接性。测验结果表示的是考生在测验活动中的外在表现，它并没有直接体现能力的实质水平。由于人们使用数量去刻画考生外在表现与内部心理特性与心理水平的关系，而这种关系属于不确定的相关关系，只能用概率模型加以描述，因此，基于这种相关关系所获得的测验结果是否有效、准确地反映出学生的能力状况，需要进行相关的统计检验。

2. 度量单位与考生位置的相对性

由于人的外显行为体现为连续的过程，不存在绝对的零点（即外在表现为 0 时，不存在相应内在心理特性全无的情况），因此教育测验的关键在于将被试放在某一行为序列上，找到被试的相对位置。这种相对性体现为两个层面，一是行为序列的相对性。例如，某个考生的数学能力在小学生数学能力群体中的位置与在中学生数学能力群体中的位置是截然不同的。二是考生位置的相对性。考生在某次教育测验中的表现既有必然因素又有偶然因素的影响，不一定恰好反映出其实际能力，因此在某次教育测验中位于某个群体的某一位置具有相对性，不能简单地断言这就是该生的实际水平。

3. 对测验施测主体水平的依赖性

教育测验的对象主要是学生的精神属性，它需要通过分析学生的外在行为、言语活动等来加以推论、解释与揭示，所采用的测验卷并不像物理量（如长度、体积、质量等）的测量参照物（如米尺、量杯、天平等）那样界限明确、状态固定，测验卷的质量依赖于命制者的专业水平，而根据考生的答卷情况判别考生能力的过程与量化结果也严重依赖于教师的经验水平。因此，教育测验的量化水平与结果运用的范围都是有限的，不应过度追求教育测验的准确性与代表性。

4. 测验活动中施测主体与考生的交往性

教育测验中，测验的施测主体与考生都是人，人都具有主观能

动性，对测验目的、测验内容、测验方式与方法等都会进行价值判断。因此，测验前，施测主体（主要是教师）与考生需要就一些关键性的问题达成共识，避免抄袭、不理解、不配合等不良现象的发生。

二、教育测验的功能

在学校教育中，教育测验是教学过程不可缺少的环节，它可以在教学的不同阶段进行，其功能主要体现在教学诊断、教学评价、促进学生学习等三大方面。[①]

1. 教学诊断方面

根据测验卷的命制、测验实施、测验后的试卷分析，教学测验可以向教师提供多种信息，作为教师诊断学生学习状况、了解学生的能力水平、熟悉测验卷的命题技巧等的参考。

（1）了解学生知识储备的起点

不同阶段的教学测验，功能不同。在新学段、新学年的教学前，实施摸底测验，目的是为了评估学生已有的知识储备、技能掌握水平、能力发展情况。根据摸底测验的信息反馈，教师可以选择适当的教学深度与难度、制定适当的教学进度，这样就可以有的放矢地开展有效教学，因材施教。

（2）检查学生知识与技能掌握的状况

阶段教学完成后的教学测验，主要目的是检查学生是否掌握本阶段的教学内容。对测验后的学生答题情况分析，可以向教师提供丰富的学生学习状况信息，教师既可以较为全面地了解学生掌握知识的状况与水平，又可以从中查清学生在认知结构上存在的缺陷与不足，从而为实施补救教学提供详实的参考。

（3）熟悉试题的特征与功能

教育测验质量的好坏很大程度上依赖于测验卷的命制质量。

---

① 余民宁著，教育测验与评量——成就测验与教学评量（第二版），台北：心理出版社，2004年10月，25—28页。

测验后的试卷分析可以向教师提供有关试卷与试题的各种技术指标值(如:信度、效度、难度、区分度等),如果再结合教学内容与考查内容分析、学生答卷分析等进行研究,就可以鉴定试题的好与坏,认清不良试题,并进行修改或删除。教师如果在这方面经常性地留心学习,就可以掌握命题技巧、提高命题水平。

2. 教学评价方面

教育测验的结果除了用来诊断问题外,也广泛地作为对教师任教、学生学习进行问责的依据,并促使教师与学生对自己的职责承担起责任。

(1) 鉴定学生的学习成果

教育测验的基本作用之一是对学生的学习成果进行等第评定,用测验成绩作为学生相关学科学业成就的指标。这种学习成果鉴定既可以作为学生在校学习效果的量化指标,也可以作为其他教育研究用途的指标。

(2) 评定教学目标的达成状况

教学是教与学的共同体,教育测验在评定学生学习成果的同时,也评定了教师教学的效果,即教学目标的达成状况。根据测验后的结果与分析,教师可以知道:目前的教学状况是否达到了预设的教学目标?如果没有达到,则距离目标还有多远?造成这种情况的原因是什么?是教材问题、教法问题、学法问题,还是其他?是否需要进行补救教学?是否需要再进行一次测验重新加以评定?等等。

(3) 制定后继教学方案的参考

教育测验的结果除了评定学生的学、教师的教,还促进教师认真反思自己的教学得失,如教学过程中教学内容的深浅、难易、宽窄等的把握是否恰当,教学策略与方法是否合适,对学生学法的指导与督促是否到位,对考试技巧的关注与指导是否适度,与其他教师的沟通是否及时,等等。对这些问题的思考为制定后继内容教学方案提供了重要参考。

3. 促进学生学习方面

教育测验不仅能够帮助教师诊断教学问题、改进教学活动,也

能够帮助学生积极开展学习，促进个人成长。

（1）激发学生的学习动机

一份编制良好的教育测验，可以向学生提供多方位的信息，如短期学习目标的检核、主要学习方法与技巧的引导、学习成果的反馈，等等。学生参加测验后，能够明确感受自己的进步，清晰看到自己是否达到学习目标，这样就可以激发学生的学习动机，每个阶段都自主地向教师制定的教学目标靠近。

（2）引导学生系统复习与钻研知识

教育测验的目的之一是促进学生更好地开展学习。学生在参加测验之前，需要对阶段的学习内容进行系统复习；参加测验时，需要集中精力对自己的记忆力、理解力、推理力、迁移力等进行全方位的调动，尤其在解答较为复杂、深奥的难题时，需要对自己的潜能进行深层次开发。教育测验是对教学活动的有力补充。

（3）帮助学生客观认识自我

教育测验后，学生通过总结与反思，可以发现自己学习方法、学习习惯等方面的优势与不足，洞悉自己在知识学习方面的长处与短处、缺陷与漏洞、困难与疑惑，明确自己在概念、法则、命题、性质与定理、技能与方法等方面的学习状况，以促进自我了解、自我认可，在后续的学习中开展针对性学习，形成最佳的学习决策。

三、常模参照测验与标准参照测验

测验是测量的一种。凡是测量都有参照系，如，测量山的高度总要对照海平面；教育测验的分数也必须与一定的参照物或参照系进行比较，其意义才能说清楚。按照解释测验分数的参照体系的不同，教育测验可以分为常模参照测验和标准参照测验两种。

1. 常模参照测验

参照某个特定群体分数的常模来解释个别测验分数在团体中所处相对位置的测验称为常模参照测验，常模参照测验分数是一种相对评分分数，所测考生分数的意义是由他与其他考生水平相互比

较的关系来确定的，它突出的是个别差异。因此，在进行分数解释前，常模的选择是关键。

常模是解释分数意义的参照体系，是对个别测验分数做出相对意义评定的实际依据。教育测验中经常使用的是组内常模，它主要分为百分等级常模与标准分数常模两类。

(1) 百分等级常模

百分等级常模就是某个特定考生群体(即常模样组)所测特定水平百分制分数的水平分布状态。在实际操作时，人们只需要将某一考生的测验分数与常模样组的整个分布状态进行对比，就可以确定该考生的水平在常模样组中的相对位置，从而确定其优劣。

一般而言，不同测验上的原始分数不能直接进行比较，因为各个测验的满分值、难度等不同。但如果不同测验在同一个常模样组上建立了常模，就可以把原始分数对应的百分等级值(称为导出分数)找出来，通过导出分数的比较来确定原始分数的优劣。

百分等级分数与原始分数的基本转换公式如下：

$$P_R = \frac{100(N-R)+50}{N}。 \qquad (1.2)$$

其中，P_R 为百分等级分数，N 为常模样组的个体数，R 为常模样组中全体原始分数按照由大到小的顺序排列后，某一原始分数所占的名次。

另外，百分等级仅仅具有顺序性，不具有等单位性(如 90 分和 89 分间的 1 分与 59 分和 60 分间的 1 分不等值)，因此，百分等级常模的取值只具有可比性，不具有可加性。

(2) 标准分数常模

标准分数常模就是某个特定被试群体(即常模样组)所测特定水平标准分数的水平分布状态，其中标准分数与测验分数的转换公式如下：

$$Z = \frac{X-X}{S}。 \qquad (1.3)$$

其中，X 是某一测验分数，$\overline{X}$ 是常模样组的测验分数平均数，S 是常模样组的测验分数的标准差，Z 即相应测验分数的标准分数。

由于无论什么测验，原始分数转化为标准 Z 分数后，其平均值均为0，标准差均为1，因此，只要测验都是向同一个群体施测，这些测验的 Z 分数就有相同的参照点与单位。因此，标准 Z 分数不仅具有可比性，也具有可加性。

实际应用中，标准 Z 分数可能出现负值与小数，容易引起误解，因此，在具体使用时，往往对标准 Z 分数进行线性变换，例如，广东高考曾使用 $T = 100Z + 500$ 来公布考生的高考单科成绩与总分成绩。

常模参照测验的目的旨在区别不同考生间的不同成就水平，并给参加测验的考生评定学业成就等级。一般地，学校使用的学期测验、模拟测验、学科竞赛，以及多数与就业或升学选拔有关的测验（如联考、高考等），都属于常模参照测验。

2. 标准参照测验

标准参照测验是指参考国家制定的课程标准或学校与教师在教学前制订的标准来解释个别测验分数是否达到这项既定标准的一种测验。这里需要明确以下几个要点。①

（1）参照的标准应该给出良好的界定，容易量化；

（2）编制的测验卷对参照标准的各个项目具有良好的典型性与代表性；

（3）参照标准对被试的表现做出的是绝对评分；

（4）测验结果应能够确切地反映被试在参照标准领域中实际掌握了什么，掌握到什么程度，还有什么没有掌握。

标准参照测验的目的旨在了解学生已经学会了什么，是否达到某一阶段学校或教师所期望的成就水准，而不是将学生与他人进行

---

① 漆书青著，现代测量理论在考试中的应用，武汉：华中师范大学出版社，2006年12月，124页。

比较。通常，学校中的平时小考、课堂测验、单元测验等，以及毕业资格考试、学科结业考试、高中会考等，都属于标准参照测验。

第三节　教育统计简介

统计，最早出于拉丁语 Statu，原意是指各种现象的状况或状态，由拉丁语词根组成意大利语 Stato，是国家的概念与有关各国结构和国情方面知识的总称。[①] 现代统计包括统计资料的收集、整理、分析以及据此进行有关判断与决策的全过程。

一、教育统计的含义

教育统计是统计的一个分支，是把数理统计的理论与方法应用到教育领域，侧重于从数量角度研究教育现象、教育规律的一门应用统计学。[②] 教育统计的主要任务是研究如何收集、整理由教育测验与教育调查所获得的大量数据资料，运用多种统计方法进行分析与推断，为做出正确的价值判断提供依据。

在教育领域中，人们借助统计方法对数据加以组织与概括，以揭示数据之间隐含的规律；通过统计分析，刻画教育现象间的相互关系；借助对样本的研究，对相应总体进行统计推断；研究教育现象中的各种差异，对差异进行显著性检验，探索差异的有效性，并力求正确反映差异间的关系；对教育现象之间的联系与变化进行深入分析，试图判别影响教育现象变化的因素，等等。统计方法的使用极大地提高了教育研究的科学性与规范性。

二、教育统计的内容

从应用的角度来分，教育统计主要包括描述统计、推断统计两

---

① 中国社会经济统计百科全书，湖北教育出版社，1994 年版，32 页。

② 王景英主编，教育统计学（第 2 版），北京：高等教育出版社，2006 年 10 月，3 页。

方面的内容。

1. 描述统计

描述统计主要是对已经获得的数据进行整理、归纳、概括，使得数据的数量分布特征或变化趋势得以清晰、明确显现。

描述统计的作用就是提供描述、概括数据的具体方法，它包括两个方面。一是绘制统计图表，如：频数分布表、频数分布直方图、折线图、茎叶图等，以直观形象的方式反映出统计数据的分布特征。二是计算统计量数，如集中量数（平均数、中位数、众数等）、差异量数（标准差、极差、方差等）、相关系数、对比率等，以简约、概括的数学语言反映数据的特定特征。

描述统计使得人们可以采用一组公认的、标准的描述数据，客观地反映出数据的特征，并消除数据解释过程中的主观偏见，它是一种让人读懂数据的工具。

2. 推断统计

推断统计的基础是概率论，它是在一定概率意义下，通过已知部分的数量特征信息，对未知总体的数量特征与数量关系进行推测与估计的统计分析方法。当人们完成对一组数据的推断后，应以概率而非绝对真理的方式陈述结论，如“有 90%的把握认为学生的入学成绩与高考成绩呈正相关”。

统计推断的内容包括两个部分。一是总体参数估计，即采用“点估计”或“区间估计”的方法，用样本的数字特征对总体的数字特征进行估计。二是假设检验，即首先对总体的数字参数或分布状态提出一个假定性判断，在这一前提下，根据样本提供的信息，采用“参数检验”、“非参数检验”的方法，在一定概率意义下，对假设前提做出接受或拒绝的决策。

推断统计在已知与未知间、有限与无限之间架起一座桥梁，使得人们研究问题的思路、方法和研究领域等都得到了极大的拓广。

三、教育测验与教育统计的关系

教育测验与教育统计都借助数学方法开展研究。教育测验是

依据一定的法则，用数字对教育过程或教育效果加以确定的过程。由于教育测验对象不是实体存在的，是人类的心理特质，因此，教育测验具有测验方法和结果表示的间接性、度量单位与被试位置的相对性、对测验施测主体水平的依赖性、测验活动中施测主体与被试的交往性。教育统计同样以数据作为研究对象，以数学方法作为研究手段，对教育教学现象进行分析，力图揭示其中隐含的教育规律。

教育测验与教育统计是了解教学状况、开展教学活动、进行教学评估的有效工具，二者总是共同使用、共同发挥作用。其中，测验与测量是统计的前提，有了测验与测量的数据，统计才可以进行；反过来，统计又是测验与测量的基础，不借助统计方法，测验与测量的数据无法进行整理、概括与分析。

第二章 试卷质量的基本分析

在学校教学工作中，教师必须运用教育测量理论设计各种测验，对教与学的情况做出评估与决策。评估与决策包括两个方面，一方面，通过测验，教师可以了解学生的学习情况，并借此了解教学效果，改进教学方法，提高教学质量；另一方面，学生可以通过测验，了解自己对学习内容的掌握情况，有针对性地总结学习方法，提高学习效益。

试卷是测验运行的实际载体，试卷命题质量的优劣，直接关系到根据测验进行评估的有效性与决策的正确性，对试卷进行科学、客观的评价分析，对优化学与教的内容、改革学与教的方式、把握学与教的重点、提高学与教的质量，以及加大对学校测验的管理力度，均具有重要意义。

一般地，对试卷命制质量的分析，往往放在学校教学测验结束后进行，通常从定性、定量两方面开展。定性分析时，主要考虑：试题测查的内容要求实际上能否达到原定目标；试题类型编制原则运用得如何，实际编拟技能发挥得如何，包括情境设置、问题提出、作答指导、词语表达和图形符号等方面；试题间关系的处理是否适当；评分标准是否正确、科学、合理、明晰，等等。定量分析主要是计算试卷与试题技术质量的指标或参数，包括难度、区分度、猜测概率、选择题干扰项效率等。

根据经典真分数理论，分析一份测验试卷的质量，既需要考虑测验结果的稳定性与一致性，即测量结果是否真实、客观地反映考

生的实际水平；又要注重测验结果是否准确有效，即测量结果是否能够反映预期的测验目的；还要兼顾试题的难易程度与考生知识与能力水平是否相匹配，是否能够将学习能力不同的考生区分开来。这涉及四个基本质量指标：信度、效度、难度和区分度。其中，信度与效度主要针对整份测验卷而言，难度与区分度主要针对测验试题。

第一节　测验信度

对整份测验试卷而言，测量的结果是否真实、客观地反映了考生的实际水平，即测验可信与否，在多大程度上可信，是首先需要考虑的问题，这就是测验信度。

一、信度概念

测验信度指的是测量结果的稳定性或可靠的程度，即测验得到的结果（实测分数）与考生实际水平（真分数）间的一致性程度。由于种种原因，实测分数一般并不等于考生能力的真分数，两者之间的差异值称为测量误差。显然，测量误差越小，测验信度就越高。

根据经典真分数理论，测验信度可以从真分数模型、依据平行测验构造的真分数等值模型等角度理解。

1. 信度指数

根据真分数理论模型 $x = t + e$，实测分数 x 不仅受真分数 t 的影响，还受误差分数 e 的影响。由于真分数 t 与误差分数 e 彼此独立，实测分数 x 与真分数 t 的相关性也决定着误差分数的大小。如果 $\rho(x, t) = 1$，则表明误差完全被控制，实测分数 x 能够毫无偏差地传达真分数 t 的信息；如果 $\rho(x, t) = 0$，则表明实测分数 x 与真分数 t 毫不相干，实测分数的差异反映的全部是随机误差的影响，说明测验结果毫无意义。因此，$\rho(x, t)$ 可以反映对测验误差控制能力

的大小，它被称为信度指数。依据真分数理论假设可以推导出

$$\rho(x,t)=\frac{\sigma_t}{\sigma_x}。\tag{2.1}$$

其中，σ_t 是真分数的标准差，σ_x 是实测分数的标准差。

由于理论上真分数不可直接测量，公式 2.1 无法用于实际计算，只有理论分析的价值。

2. 信度系数

根据真分数理论，一组平行测验测量的是同一个考生总体的真分数，若测验控制误差的能力强，则无论用平行测验中哪一个测验上的实测分数去估计真分数都不会有过大偏离，所以用一个平行测验 x_1 上的实测分数去估计另一个测验 x_2 的实测分数也应该很准确。也就是说，当测验误差控制能力强时，两平行测验上实测分数的相关性就高。由于一组平行测验中任意两个测验的相关系数都等于其他任何两测验的相关系数，因此，我们称平行测验上两组测验实测分数间的相关系数 $\rho(x_1,x_2)=r_{x_1x_2}$ 为测验的信度系数。容易推导得出：

$$r_{x_1x_2}=\rho(x_1,x_2)=\frac{\sigma_t^2}{\sigma_x^2}=[\rho(x,t)]^2。\tag{2.2}$$

其中，σ_t^2 是真分数的方差，σ_x^2 是实测分数的方差，$\rho(x,t)$是测验的信度指数。

公式 2.2 中，虽然真分数的方差无法计算，但是两个平行测验的实测分数却是可以得到的，这样就给解决信度计算问题开辟了实际操作的途径。

在数学测验研究中，我们往往只探讨信度系数的计算问题。本书以下没有特别说明时，信度都是指信度系数。

二、测验信度的估算方法

理论上，我们可以编制出两份平行测验卷，然后让同一批考生

使用这两份测验卷，再根据得到的两组实测分数计算两份测验卷的相关系数，从而估计测验信度。然而，实际工作中，一方面，教师和学生没有那么多的精力和时间参加重复的测验；另一方面，在实际操作时，也很难编制出严格意义上的平行测验卷，很难保证两份试卷测量的是同一种数学能力，测验的题型、方式完全相同，内容覆盖完全相同，难度完全等值。

在具体应用中，人们对估算方法进行了改进，希望根据一次测验来估计测验信度，通过这种方法估算得到的测验信度都称为测验信度的内部一致性系数。估算测验信度内部一致性系数的方法主要有两种。

1. *分半法*

顾名思义，分半法就是将测验施测于一组考生，然后将测验人为地分成两个平行部分，通过比较这组考生在这两个部分实测分数间的相关性，来估计测验信度，这样得到的测验信度也叫做分半信度。

分半法的核心在于如何将测验卷分半。适合数学测验使用的分半方法有两种。第一种是奇偶题目分半，即将奇数题（第 1、3、…题）组成一个部分，偶数题（第 2、4、…题）组成剩余部分；这种分半法可以保证两个分测验卷都包容了原测验的开头、中间、结尾的同等数量的题目，因而平衡了很多干扰效应。第二种是将测验卷分成若干个内容块，再将各内容块的题目奇偶分半，所有的奇数题和所有的偶数题各组成一个分测验。具体命题时，可以考虑把两种方法结合起来组构测验卷。

根据分半法得到分测验 x_1 与分测验 x_2，计算得到的相关系数 $r_{x_1x_2}$ 只能代表半个测验的信度，它并不是整个测验的信度，需要进行矫正，矫正公式如下：

$$r_{xx}=\frac{2r_{x_1x_2}}{1+r_{x_1x_2}}。\tag{2.3}$$

其中，r_{xx} 为整份测验卷的信度，$r_{x_1x_2}$ 为两个分测验卷间的相关

系数。

分半信度的误差主要来源于测验的分半过程。用不同方式对测验进行分半,所得的分半信度值也会不同。

2. 同质性法

同质性法是分半法的拓展,分半法是将一份测验卷分成两半进行估算,同质性法是将一份测验卷分成 n 个平行的部分,求这 n 个平行的部分间的一致性程度。计算公式为

$$r_{xx}=\frac{nr_{x_1x_2}}{1+(n-1)r_{x_1x_2}}。\qquad(2.4)$$

其中,r_{xx} 为整份测验卷的信度,$r_{x_1x_2}$ 为 n 个平行部分任意两个间的相关系数。

当 $n=2$ 时,即为公式 2.3。

当把测验卷中的每道试题看成彼此平行时,公式 2.4 可以变形为

$$r_{xx}=\frac{n}{n-1}\left(1-\frac{\sum_{i=1}^{n}p_iq_i}{\sigma_x^2}\right)。\qquad(2.5)$$

这里,n 是测验卷中的试题总数,p_i 与 q_i 分别是第 i 道试题上的答对率与答错率($q_i=1-p_i$),σ_x^2 是测验总分的方差,这就是有名的库德—理查森公式(KR-20)。该公式仅仅适用于估算由"对、错"分为两级评分的选择题组成测验卷的信度。

当测验卷的题型有填空题、选择题、解答题等多种形式,每道试题的满分也不一定相同时,公式 2.4 可以变形为

$$\alpha=r_{xx}=\frac{n}{n-1}\left(1-\frac{\sum_{i=1}^{n}\sigma_{x_i}^2}{\sigma_x^2}\right)。\qquad(2.6)$$

这里,n 是测验的试题总数,$\sigma_{x_i}^2$ 是每道试题上实测分数的方差,

σ_T^2 是测验总分的方差，这便是有名的科隆巴赫 α 系数公式。

需要注意的是，测验信度不仅与测验试题总数相关，而且与考生群体的大小也有很大的关系，计算时，需要特别交代信度计算的背景。

三、影响测验信度的因素

根据经典真分数理论，测验信度主要反映的是测验中控制误差能力的强弱，如果明确了误差来源，就能够有目的地加强对误差的控制，从而提高测验信度。因此，有必要对测验误差的来源进行探讨，下面分别从测验卷与测试题、考生因素、施测环境、评分标准与评分过程等四个方面进行分析。

1. 测验卷与测试题

测验卷与测试题本身的一些因素会直接产生测量误差。例如，数学测验卷中都包括客观题与主观题两大类，其中客观题中采用选择题、判断题、填空题等多种形式，这些题型由于直接根据结论的对错判断考生对相应知识点的掌握情况，缺少对考生思维过程的分析，考生在作答时答案的获得又具有一定的猜测性，这就可能导致实测分数与考生真实能力之间不一致，从而影响测验信度。

又如，测验卷中试题难度直接影响考生成绩。如果试题难度过大，考生凭借掌握的知识不能顺利解决，那么考生就会通过猜测来给出解答以获取一定的分数，这时实测分数更多反映的是测验误差，测验信度很低。反过来，若试题越容易，考生得分就越高，就越不能看出考生实际学习能力的差异，因此测验信度也就越低。

另外，试题的取样是否具有代表性、测验卷规定的作答时限是否足够、测验卷的试题总数多少、测验卷中试题知识点分布状况、测验卷考查能力的针对性等因素都会影响考生真实水平的发挥。

2. 考生因素

考生状态是影响测验信度中最难控制的因素。首先，考生的应试技巧与稳定的反应倾向直接影响考生真实水平的发挥。应试技巧在某种程度上可以有效弥补知识与技能方面的不足。其次，考生

的应试动机、情绪的紧张焦虑状态也会影响其作答结果。应试动机会影响到考生的作答速度、注意力、耐久性、反应速度等，适度的焦虑水平将提高考生的兴奋性，增强其注意力；而测验的时间限制、难度水平等决定着考生的焦虑水平，因此，种种因素都可能对测验结果产生积极或消极影响。另外，考生对数学测验任务的理解、在所测知识技能上的熟练水平、在记忆与注意力上的波动等都可能对考生的测验操作水平产生不稳定的影响。最后，考生的健康状况、疲劳与否等因素也会影响其真实水平的发挥。

3. **施测环境**

施测环境包括物理环境、非物理的外界环境两类。物理环境包括：施测教室的光线、噪音、通风、温度等，这些都可能对考生的情绪、应试状态以至真实水平的发挥产生正面或负面的影响。非物理的外界环境包括：有人作弊、试卷发错、临时发现试卷印刷不清等种种意外的干扰与突变，导致考生分心，影响考生作答。

4. **评分标准与评分过程**

由于数学主观题的解答方法往往不止一种，如何使得不同解法之间的评分标准保持等价，如何确定不同步骤之间的赋分，都具有很强的经验性。其次，不同教师对主观题的同一个评分标准的理解也会有差异，在评判考生的具体作答时也会有差异。另外，同一个教师因为心境、疲劳等因素的影响，评卷过程的前后标准也会有差异，这些都可能导致评分误差的产生。

由上述分析可知，改进测验信度指标可以从以下几方面进行：适当增加试题总数、保证一定的试题难度、减少试题的猜测度、制定合理的评分标准、保证试题考查内容的针对性、控制外在环境减少无关干扰等。

四、测验信度的定性评价指标

对一份测验卷的信度评价，除了计算一个具体数值并交代该数值反映的实际意义外，从定性的角度展开分析更有助于测验卷命制

质量的改进。一般地，在施测外在环境已经确定的情况下，定性评价测验信度主要围绕测验卷、测验作答过程与测验评分过程进行，因此常见的评价指标有如下四个。

1. 试题考查目标的一致性程度

学校教学过程中的测验目的很明确，即为教与学服务。如果测验卷中的试题围绕着某阶段的教学内容命制，且试题的背景为学生所熟悉、不存在理解障碍，试题涉及的知识技能针对性强，知识技能跨越度合适、不存在大面积遗忘现象，那么实测分数就能够较好地反映出考生阶段性知识技能的掌握情况，测验比较可信。

2. 试题呈现的规范性

在数学测验卷中，试题呈现的规范性包括以下四个方面：文字表述准确，不会导致理解歧义；数学符号呈现规范，使用合理；几何图形、统计图表等呈现规范，标注准确，线条美观大方；不存在试题跨页现象，不会导致考生无意错漏等。当测验满足了试题呈现规范性要求后，就可以尽量减少试题因素造成的测验误差，提高测验信度。

3. 试题作答的猜测度

由于试题猜测度过高会影响对考生真实水平的判断，因此减小试题正确作答的猜测度，就可以提高测验信度。在分析选择题、填空题的作答时，应研究正确作答的猜测度的大小，尽量减小正确作答的猜测度。

4. 评分标准的合理性

教学过程中用于检查教学效果的测验大部分由教师决定评分标准，而评分标准是否合理直接关系到实测分数与考生真实水平之间的一致性。在具体研究评分标准时，主要看同一试题不同解法评分标准的等价性、同一试题解答分步赋分的合理性、试题难度与分值多少的匹配性等因素。

根据上述分析，在评价测验信度时，可以先由各个评卷教师填写表 2-1，然后再将评卷教师的意见汇总形成一份测验卷的总评价表。填表时，其中正面典型、负面典型主要举出各个指标中做得好

或不好的案例，并简要说明好或不好的原因，以便以后改进。

表2 1 测验信度定性评价表

评价指标	评价要点	评定等级				正面典型	负面典型
		优	良	中	差		
1. 试题考查目标的一致性程度	(1) 卷中试题围绕着考查目标命制						
	(2) 试题的背景为学生所熟悉，不存在理解障碍						
	(3) 试题涉及的知识技能针对性强						
	(4) 知识技能跨越度合适，考生不存在大面积遗忘现象						
2. 试题呈现的规范性	(1) 文字表述准确，不会导致理解歧义						
	(2) 数学符号呈现规范，使用合理						
	(3) 几何图形、统计图表等呈现规范，标注准确，线条美观大方						
	(4) 不存在试题跨页现象，不会导致考生无意错漏						
3. 试题作答的猜测度	(1) 选择题选择支的迷惑性好						
	(2) 主观题的难易程度合理						
4. 评分标准的合理性	(1) 同一试题不同解法评分标准的等价性						
	(2) 同一试题分步解答赋分的合理性						
	(3) 试题难度与分值多少的匹配性						

五、测验信度的定量评价标准

测验信度值要达到多高才好？这取决于测验的使用目的与测验的类型。在教育测验中，一般以科隆巴赫 α 信度系数作为测验信度的下限，来评价测验质量。根据纳讷莱(1967)的研究，α 信度系数的不同范围，反映测验的不同问题，如表 2－2。[1]

表2－2　科隆巴赫 α 信度系数评价标准

α 值	评　价
[0.9，1)	信度很好，达到最好的标准化考试水平。
[0.8，0.9)	对学校考试而言，非常好。
[0.7，0.8)	对学校测试而言，大部分试题很好，可能少数试题需要改进。
[0.6，0.7)	信度稍低，需要补充其他测验以确定分数或等次。部分试题需要改进。
[0.5，0.6)	信度低。建议对试卷进行修改(如果试题数多于 10 道)。需要补充其他考试来可靠地确定分数或等次。
(0，0.5)	信度差。考试基本无效，需要修改。

然而，结合教学过程进行的常规数学章节测验、单元测验与学期测验基本上属于标准参照性测验，突出的是教学目标与课程标准要求，目的是为了告诉考生应该学习什么、已经掌握了什么、掌握到什么程度、到底还存在什么差距，关注的是教学目标的全面达成与实现。在平时的小测验中题目的同质性要求并不是主要关注的，因此，可能出现测验信度的内部一致性系数较低，但测验结果的可靠性不低的情况。

需要注意的是，在估计一份测验卷的测验信度时，考生总体异质性高，即所测真实能力差异大，则实测分数分布范围就广，测验信度值相应就高。反之，若考生总体同质性高，那么实测分数彼此很

① 雷新勇，考试数据的统计分析和解释，上海：华东师范大学出版社，2007，246 页。

接近，分数分布就很狭小，对应的测验信度值就偏小。

第二节　测验效度

测验信度高低代表的是测验分数的稳定性，它是评估测验质量的一个重要指标。但是，测验信度高不一定就可以断言这是一个好测验。例如，用直尺多次度量某个物体的长度，所得结果的误差非常小，但是如果该直尺的刻度本身有错误，那么所度量到的物体的长度总是以一定的规律偏离物体的真正长度，那么这种度量显然是无效的。又如，如果用直尺去衡量物体的重量，那么显然无法达到度量目的。因此，评估测验质量时，还需要使用另一个更为重要的指标：测验效度。

一、效度概念

效度，就是测验测到计划要测的东西的程度。例如，如果一个测验是去评估考生数学推理能力的，那么需要确定的是考生在测验过程中应该是对给定的材料进行推理，而不是在生搬硬套公式。如果测验结果完全反映出编制测验时希望测到的数学推理能力，那么测验效度就很高；如果实际上只测出了一部分数学推理能力，其余测出的是考生的记忆力，那么测验效度就不够高。因此，效度是测验中最重要的质量指标。

由于测验要测的对象是考生的心理特质（知识与技能掌握情况、空间想象能力、推理能力等），所编制的测验是否真正测查到它，即测验是否有效、有效的程度如何，并不能由人们的主观愿望、看法、经验等简单决定，而只能依靠客观事实和实际证据进行验证。验证测验效度时需要明确以下几个问题。

1. 测验是否正确有效，首先取决于测验目标

任何测验都是为测验目的与功能服务的。为了支持测验目的，

在命制测验卷时需要根据测验目的界定应予评鉴的知识、技能、能力、过程与特征，并说明它与其他测验目的的异同。例如，某个阶段单元测验目的是为了检验学生是否做好学习下一阶段知识的准备，那么命制测验卷时需要考虑：(1)学习下一阶段知识必备的技能是什么？(2)与这些必备技能相对应的测验内容领域是什么？(3)如何选择测验试题才能代表测验内容领域？(4)测验分数是否会受无关变量的过多影响？(5)测验得分高的考生是否在下一阶段学习中能够比测验得分低的考生学得更好？等等。同样地，获取实测分数后，也要从这些方面去验证测验目标实现了多少，进而去判断测验效度的高低。

由于不同测验服务于不同的测验目的，因此，对于某一目的而言效度很高的测验，对于其他目的而言，测验效度可能很低，这一点需要注意。

2. 测验是否有效，关键取决于如何使用测验结果

测验最终是为了解决一定的问题。因此，测验是否有效，关键要看测验结果如何使用，即能否对实测分数进行合理、正确的解释，以及依据这些解释能否做出有效、可行的决策。

如果说，人们根据测验目的精心制作了测验工具，测验工具的优劣根据测验目的来评鉴，那么，工具能否得到良好使用则与测验实施过程、施测结果紧密相连。例如，根据阶段单元测验目的是“为了检验学生是否做好学习下一阶段知识的准备”命制出单元测验卷，考生完成单元测验卷后，教师没有对实测分数进行针对性的解释(如，得分在多大程度上反映出考生对关键知识技能的掌握情况)，也没有判定这一分数能否作为开始下一阶段知识教学的依据(如，可能学生知识储备不足，不能进入下阶段教学；或者学生掌握情况非常好，下阶段教学可以适当简化某些内容；等等)，就按照教学计划推进教学进度，那么这次单元测验的效度就不够理想。

3. 测验效度有高低之分，无有无之别

在学校教学中，由于测验总是为一定的教学目的服务，也要解

决一定的教学问题，因此测验总是能够或多或少地反映出考生的某些心理特质。在具体评估一项测验的效度时，人们往往根据收集的信息从效度高或效度低的角度加以验证，而不是直接断言测验有效或无效，这一点也需要注意。

二、效度与信度的关系

在经典真分数理论中，实测分数的方差 σ_x^2 可以分解为真分数方差 σ_t^2 与测量误差方差 σ_e^2 之和，即 $\sigma_x^2 = \sigma_t^2 + \sigma_e^2$。测验信度定义为 $r_{xx} = \sigma_t^2/\sigma_x^2$。

事实上，测量误差包括随机误差和系统误差两部分，由于系统误差与所测特质无关且较为稳定，因此经典真分数理论在研究信度时把系统误差放在真分数中不做进一步分解。

在效度研究中，系统误差属于重点研究对象，因此，人们把真分数方差 σ_t^2 进行再分解，得到公式 2.7

$$\sigma_t^2 = \sigma_V^2 + \sigma_I^2, \tag{2.7}$$

即

$$\sigma_x^2 = \sigma_V^2 + \sigma_I^2 + \sigma_e^2。\tag{2.8}$$

其中，σ_V^2 表示考生真实能力水平的方差，σ_I^2 表示测验中的系统误差方差。

效度的估算公式定义如下：

$$r_{xy} = \frac{\sigma_V^2}{\sigma_x^2}。\tag{2.9}$$

其中 r_{xy} 表示测验的效度，σ_x^2 表示考生实测分数方差。

由于考生的真实能力水平是测验希望测到的，无法直接获取，因此根据公式 2.9 无法计算出测验的效度值，公式 2.9 只是具有理论研究价值。

根据公式 2.7 与 2.9，可以得出如下的等量关系：

$$r_{xy} = \frac{\sigma_V^2}{\sigma_x^2} = \frac{\sigma_t^2 - \sigma_I^2}{\sigma_x^2} = r_{xx} - \frac{\sigma_I^2}{\sigma_x^2}。\tag{2.10}$$

根据公式2.10,对于同一份测验,可以得出以下结论:

1. 效度值总是小于信度值。

2. 如果测验信度低,则测验效度肯定低。

3. 如果测验信度高,测验效度不一定高,此时,效度的高低取决于系统误差的大小,系统误差大,则效度低,系统误差小,则效度高。

4. 如果测验效度高,则测验信度一定高。

与信度分析一样,在具体分析一份测验时,也需要从定性与定量两种角度来分析测验效度。

三、效度分类与评估

将效度分为内容效度、结构效度、效标关联效度是经典的、公认的分类方法,至今仍在效度研究领域占据重要地位,在中小学教学领域的应用也很广泛,下面逐一介绍。

1. 内容效度

(1) 内容效度的概念

内容效度是指测验的题目在多大程度上代表了所欲测试领域的整个内容。例如,如果要测试学生简单的算术计算能力,那么使用一份由30道四则运算题组成的测验卷比使用一份由10道算术应用题组成的测验卷更有效,也就是说前一份测验卷的内容效度更好,测验内容更具有代表性。

教育测验中,由于学生学习的知识很多,教学应达到的目标也很丰富,而测验不可能包罗万象,因此,只能选择一部分知识、主要的教学目标构建测验卷作为样本,去估计考生的知识技能掌握情况,并对取样是否合适做出估计。例如,从数学某单元的内容中进行取样形成课堂小测验,用学生在这个题目样本上所得的分数推测学生在相应单元学习中知识、技能与能力状况。如果试题样本的代表性好,那么测验结果就可以推广到整个欲测试领域中;否则,推广测验结果很可能产生错误。

由于学生所学的知识与技能分为了解、理解、掌握、灵活运用等

多种层次，因此，在构建测验卷时，除了考虑知识与技能的代表性外，还必须兼顾知识与技能的考查层次、考查方式、呈现形式等多种因素，这样才能保证测验内容具有较好的代表性。

(2) 内容效度的评估

目前并没有简单有效的公式用以计算内容效度值。

确定内容效度是否理想的关键是分析试题的取样是否具有较好的代表性。一般要求试卷中的每一道试题都必须有自己明确有效的考查目标，既要与试卷中其他试题相辅相成，又要为试卷中别的试题无法代替。

一般可以编制一个或几个考试知识和水平的双向细目表来检验内容效度，以下是较为常见的几种双向细目表。

表 2-3　反映测验内容与认知水平关系的双向细目表

学习内容		测验内容	认知水平				分值合计
			了解	理解	掌握	灵活运用	
概念	1.						
	2.						
法则或技能	1.						
	2.						
综合	1.						
	2.						
其他							
分值合计							100 分

在表 2-3 中，测验卷的满分为 100 分，表格主要包括三大部分：欲测内容范围（即学习内容）、测验内容、测验内容所属的认知水平（即检验教学目标的达成度）。根据该表，可以清楚地了解测验内容对欲测内容的代表程度、测验知识技能水平的分布状况，从而评估测验的内容效度的高低。

表 2-4　测验内容、认知水平与测验题型的三向细目表

题型 内容与水平			填空题					选择题					解答题					题数、分数小计
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
学习内容	概念																	
	法则																	
	技巧																	
	思想																	
认知水平	了解																	
	理解																	
	掌握																	
	灵活运用																	
每题满分值			5	5	5	5	5	5	5	5	5	[illegible]	10	10	10	10	10	100 分

在表 2-4 中，测验卷共 15 道试题，满分为 100 分，表格上要包括四部分：欲测内容范围（即学习内容）、测验内容分布、测验内容所属的认知水平（即检验教学目标的达成度）、测验题型的应用情况。该表对表 2-2 进行了改进，对试卷题型结构也进行了分析。

在实际构建双向细目表时，可以根据具体情况对上述两表进行适当增删，例如，可以把试题的难易度、各部分试题内容的比例等因素添加进去，综合进行分析。

2. 结构效度

(1) 结构效度的概念

数学测验的结构效度指测验结果与被测学生所具有的数学能力、智力等结构相符合的程度，其目的在于用心理学理论说明测验分数的意义，即用心理学观点对测验结果加以解释与探讨。

确定一个测验的结构效度，首先应从数学能力结构的心理学理论出发，导出各项关于这一数学能力结构的一些基本假设，再根据这些基本假设设计和编制试卷进行测验，得出测验结果后，由果溯因，用实验、相关、因素分析、聚类分析、路径分析等方法，来检验测验结果是否符合心理学上关于数学能力结构的理论假设。因此，结构效度不能用单一的数据指标衡量，而必须使用累积的证据进行评价。

(2) 结构效度的评估

在常规数学测验中，定性分析结构效度时，可以从分析测验问题与数学解题心理的关系、测验分数与其他考试(如同质考试、异质考试)分数的相关性等角度展开，简易做法是：制定试卷的框架结构、题型结构、能力结构、难度结构，看看是否有利于考生正常水平的发挥(如是否有人为制造的陷阱，是否兼顾不同考生的学习专长，试题的题型是否多样，评价方式是否客观……)。

统计学上，检验结构效度最常用的方法是因素统计法。用因素分析法来检验测验卷的结构效度，并有效地抽取几个共同因素，若这些共同因素与数学能力理论结构的心理特质很接近，则可以说此测验卷具有较好的结构效度，并可以此作为测验所测的特质对测验分数做出解释。因素分析时，抽取因素的方法有很多，常见的是使用主成分分析法、极大似然估计法、未加权最小二乘法、广义最小二乘法等；选取共同因子转轴的方法包括：最大变异法、相等最大法、斜交旋转等，具体使用指导可以参考相关的统计书。[①]

3. 效标关联效度

(1) 效标关联效度的概念

效标关联效度又称为经验效度、统计效度，它用测验分数和效标之间的相关系数表示测验效度的高低，效标就是检测效度的参照

---

① 例如，可以参考《SPSS统计运用实务——问卷分析与应用统计》(吴明隆编著，北京：科学出版社，2003年版)、《SPSS在教育统计中的应用》(杨晓明主编，北京：高等教育出版社，2004年版)等。

标准。效标关联效度分为同时效度和预测效度两种，同时效度是指测验与当前效标之间的关联程度，例如，用高考成绩作为效标来检验高中毕业会考成绩，计算两者之间的相关系数就是高中毕业会考的同时效度；预测效度指测验与将来效标之间的关联程度，例如，用大学一年级的成绩作为效标来检验高考成绩，计算两者之间的相关系数就是高考的预测效度。

效标是用来衡量测验效度的尺度，它不仅随着测验种类的不同而不同，而且也可能随着时间而改变，现在是一个好的效标，将来不一定就是。

(2) 效标关联效度的评估

效标关联效度的计算方法主要是通过计算各种相关系数而求得，可以采用积差相关、二列相关、点二列相关等。在常规的教育测验中，常用学生最近若干次考试的平均成绩简单地作为效标分数，计算某次测验的实测分数与效标分数之间的相关系数来估计效标关联效度，计算公式为：

$$r=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2\cdot\sum_{i=1}^{n}(Y_i-\overline{Y})^2}}。\qquad(2.11)$$

其中，X_i 为 n 个学生的考试分数，Y_i 为 n 个学生的效标分数，$\overline{X}$、$\overline{Y}$为相应分数的平均数。一般认为 r 值大于或等于 0.45 即可。

四、测验效度的定性评价指标

与信度一样，对一份测验卷的效度进行评价，从定性的角度展开分析更有助于测验卷命制质量的改进。除了上述介绍的通过双向细目表来检验测验卷的内容效度外，还可以从试题的异质性程度、试题的代表性程度、试卷结构是否有利于考生水平的发挥等角度进行分析。表 2-5 给出了一个评价量表模板供参考。

表 2-5 测验效度定性评价表

评价指标	评价要点	评定等级				正面典型	负面典型
		优	良	中	差		
1. 考查内容的有效性程度	(1) 试卷知识、技能的覆盖率						
	(2) 考查的知识技能是否涉及多个认知层次						
	(3) 知识技能考查方式的多样性						
	(4) 考查的知识技能呈现形式的多样性						
2. 试题取样的代表性程度	(1) 每道试题考查目标的有效性						
	(2) 每道试题考查功能的独特性						
	(3) 试题间考查功能的互补性						
3. 试卷整体布局的有效性程度	(1) 不同题型应用的恰当性						
	(2) 不同难度试题搭配的合理性						
4. 兼顾考生的有效性程度	(1) 是否有人为制造的陷阱						
	(2) 是否兼顾不同考生的学习专长						
	(3) 整卷的表述习惯与阅读量是否合理						

第三节 测验难度

试卷质量分析包括整卷质量分析、试题质量分析两个层面，其中试题质量分析就是根据实测资料求取试题的难度与区分度，并研

究相应的改进措施。

一、难度概念

难度是测验试题的难易程度，是对考生完成试题作答任务时所表现出来的困难程度的度量，也可以说是衡量试题对学生知识与能力水平适合程度的指标。

真分数理论中最简单最常用的试题难度指标，是直接建立在通过率基础上的。通常，试题的通过率等于该题具体的答对人数与参加测验总人数之比，按照此意，通过率越大，表示试题被正确答对的困难程度越小，难度应该越小，即试题难度应该定义为试题的“未通过率”或“失分率”。由于一道试题的得分率与失分率之和为1，两者在数学意义上是对等的，而实际工作中测验分数的统计往往是先清点答对人数或答对分数，即先求出通过率或得分率，因此，一般就用通过率（得分率）来作为试题的难度指标。

由于难度的计算总是针对特定的考生群体而言的，离开了具体的考生群体，试题就谈不上难度，因此难度具有相对性。例如，一道试题对于小学生而言可能很难，但对于中学生而言可能非常容易。在分析试题难度时，必须指明施测对象。

二、难度的计算

1. 客观题难度的计算

客观题的评分采用“对、错”方式计分，这时，试题的难度就等于答对该试题的人数与参加测验总人数的比，计算公式如下：

$$p=\frac{r}{n}。\tag{2.12}$$

其中，p 表示试题的难度指标，n 表示参加测验总人数，r 表示答对该试题的人数。

显然，难度值 p 越大，题目越容易；难度值 p 越小，题目越难。

2. 主观题难度的计算

主观题的评分采用“部分答对给部分分数，全对给满分，全错给零分”的方式计分，这时，试题的难度（通过率）就等于试题上的平均得分与该试题满分的比值，计算公式如下：

$$p = \frac{\bar{x}}{x}。 \tag{2.13}$$

其中$\bar{x}$是考生在该题上的平均得分，x为该题的满分。

3. 用极端分组法计算

当考生人数较多时，无论是客观题还是主观题，都可以采用极端分组法计算试题的难度，计算的具体步骤如下：

（1）先按照测验总分，由高到低，将考生试卷依次排序；

（2）从得分最高的一份试卷开始，依次向下选出全部试卷的27%作为高分组，计算出高分组对应的各试题的难度 p_H；

（3）从得分最低的一份试卷开始，依次向上选出全部试卷的27%作为低分组，计算出低分组对应的各试题的难度 p_L；

（4）相应试题的难度取 p_H、p_L 的平均值，即 $p = \frac{p_H + p_L}{2}$。

根据难度上述的计算公式可知，试题难度值 p 的取值范围为 $0 \leqslant p \leqslant 1$。

三、难度的评价标准

由于每个测验都是由众多的试题组成，一份测验卷中各个试题的难度如何搭配才恰当？这需要具体问题具体分析。通常，试题难度的恰当性、难度分配的恰当性取决于测验目的的性质（是标准参照还是常模参照）、所用试题的类型（是客观题还是主观题）、试题间的相关性。

对于标准参照测验，通过率越高，表明教与学的效果越好，因此，不仅每道试题难度的数值越大越好，而且试题间的难度差异也不必大。

对于常模参照测验，由于目的是要尽量把所有考生的水平差异加以区分，因此，总是希望考生的总分能彼此拉开距离，希望每个分数点上都有很强的区分度。从单个试题看，主观题（如计算题、作图、证明等）由于猜测成功的可能性很小，试题难度在 0.5 左右最为适宜，因为难度适中的试题最有利于把全部考生区别开来，过难或过易的试题不利于区分考生。客观题由于存在随机猜测答对的影响，“四选一”选择题的适宜难度为 0.7 左右，是非题的适宜难度为 0.85 左右。从整卷看，若测验中试题间的相关性很高，这时，试卷中的试题难度分布应力争宽一些，可在 0.01 至 0.99 全距间作均匀分布；若试题间的相关性很低，这时，试卷中的试题难度可以围绕 0.5 形成窄全距分布。①

对于竞赛类的测验，由于测验目的是选拔出考生中最优秀的人才，因此要求在某个分数划界点上有极强的鉴别力，以便把考生准确地分成录取与淘汰两类，因此，试题难度的数值就必须偏小。

四、影响难度的因素

一般说来影响题目难度的主要因素有：①考查知识点的多少；②考查能力的复杂程度或层次的高低；③考生对题目的熟悉程度（如本来较易的题目会因考生均未注意而造成很难，或本来较难的题目会因为考生普遍练习过而变得较容易）；④命题的技巧性（如同一个问题，可以命得容易，也可以命得较难）。

控制题目的难度，除了考虑上述因素，还可以通过其他方法来控制。在平常的教学测验中，由于教师对学生的情况比较了解，因而主要凭经验来控制难度，使之与教师的教学难度相适应。而在区级、市级大规模测验中，除了经验，还需要通过预测来掌握难度。首先由命题人员根据上述因素估计一个难度范围；然后通过小样本测试验证难度估计的准确程度，分析原因，进而提高评估能力。经过

---

① 漆书青，现代测量理论在考试中的应用，江西教育出版社，2003，89—95 页。

预测取得难度的题目可以进入题库，以备后用。[①]

五、试题难度的定性评价

测验完毕，除了计算出各题的难度值外，对难度进行定性评价也是常见的分析工作，具体可以参考表2－6进行。

表2－6　试题难度的定性分析表

题号	主要考查目标	预估难度	实测难度	其他考查方式的可能性，对教学的指导意义	考生典型解法	考生典型错误及错因分析
1						
2						
3						
……						

第四节　测验区分度

无论是考试的鉴别和选拔功能，还是诊断和信息反馈功能，都离不开测验卷的区分功能，这种区分功能用试题区分度来界定并加以测量。

一、区分度的概念

试题区分度是指测验试题鉴别考生实际能力水平高低的量度。考生的能力水平总是有高低之分，如果实际水平高的考生在测验题目上能得到高分，而实际水平低的考生只能得低分，那么该试题区分考生的能力就强；反之，就可以断定试题的区分度不理想。因此，

① 胡中锋，教育测量与评价，广州：广东高等教育出版社，2006，49—50页。

试题的区分度又被称为试题的鉴别力，它是评价试题质量、筛选试题的主要指标与依据。

根据区分度的定义，计算试题的区分度时，需要先把考生按照能力水平高低排序形成一个效标，然后再考查考生在试题上的得分情况与这个效标之间的相关性；如果相关一致性高，就是区分鉴别力强；如果相关一致性低，就是区分鉴别力差。在具体分析测验试题的区分度时，方法主要有两种：外在效标法与内部一致性法。

1. 外在效标法

即先找一个不依赖于测验成绩的、外部的、客观的标准，根据这个标准把考生按照能力高低顺序排好队，然后再看考生在测验试题上的得分，顺序是否跟前者相符。

然而，在具体实践中，这种外部的客观标准很难找到。

2. 内部一致性法

在实际操作时，一般采用内部标准，即把考生在整个测验上所得的总分当成考生的实有水平。当然，这种做法在逻辑上缺乏充分根据，原因在于，总分是否正确可靠，在分析工作尚未进行之前，无法肯定。然而，一般来说，测验都是经过一番设计的，全卷总分比起个别试题的得分来说，总是有可能更接近于考生的实际水平；另外，以总分作标准，有利于增强测验试题间的同质性，从而有利于提高整个测验的信度；而且，这种计算也可以说明每个试题应为测验目的作贡献，如果不一致，恰好就说明了该试题所测特质与测验目标不一致。

所以，在实际操作中，主要使用的是内部一致性法，从这个意义上说，区分度的实质就是各试题得分和测验总分的相关程度。

二、区分度的计算

根据测验及题目的不同计分方式，可采用不同的公式计算区分度。在数学测验中，主要采用下列几种计算方法。

1. 积差相关法

两个变量都是正态连续变量且两变量之间成线性关系时，表示

这两个变量之间的相关称为积差相关。积差相关是英国统计学家皮尔逊在 20 世纪初提出的一种计算相关的方法，故也可称为皮尔逊相关法。数学试卷中的主观题，试题得分与测验总分均为连续变量，所以一般采用积差相关公式计算试题的区分度，计算公式如下：

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}。\qquad(2.14)$$

其中，x_i 表示考生 i 在该试题上的得分，$\bar{x}$ 表示该试题所有考生得分的平均分；y_i 表示考生 i 测验的总得分，$\bar{y}$ 表示该测验卷所有考生总得分的平均分。事实上，公式 2.14 与公式 2.11 完全相同。

2. 点二列相关法

当两个变量其中一个是连续性变量，另一个是真正的二分名义变量(如，对与错)，这时，表示两个变量之间的相关，称为点二列相关。

比如，在数学测验中，客观题中选择题答对记为 5 分，答错记为 0 分，这时，选择题可以看成是二分名义变量，而测验总分看成是连续变量，这时便可采用点二列相关法计算选择题的区分度，计算公式如下：

$$r_{pb}=\frac{\overline{x_p}-\overline{x_q}}{\sigma_t}\sqrt{pq}。\qquad(2.15)$$

其中，r_{pb} 为某道试题的点二列相关系数，p 为该试题的通过率，q 为该试题的未通过率；$\overline{x_p}$ 为该试题通过者测验总分的平均值，$\overline{x_q}$ 为该试题未通过者测验总分的平均值；σ_t 为测验卷测验总分的标准差。

3. 二列相关法

当两个变量都是正态分布的连续变量，但其中一个变量被人为地划分成二分变量(如按一定分数，把测验分成合格与不合格两类)，这时，表示这两个变量之间的相关，称为二列相关。

例如，数学测验试题分数可以看成是连续变量，而测验总分被分为及格、不及格两类，可采用此法计算试题的区分度，计算公式如下：

$$r_{bi}=\frac{\overline{x_p}-\overline{x_q}}{\sigma_t}\cdot\frac{pq}{y}。\tag{2.16}$$

其中，r_{bi}为某道试题的点二列相关系数，p、q、x_p、x_q、σ_t 的含义同公式 2.15，y 为通过率 p 在正态分布中的纵线高度。

4. 极端分组法

采用极端分组法计算试题的区分度与前面计算试题难度的步骤相同，只是最后一步的计算公式不同。计算的具体步骤如下：

(1) 先按照测验总分，由高到低，将考生试卷依次排序；

(2) 从得分最高的一份试卷开始，依次向下选出全部试卷的27%作为高分组，计算出高分组对应的各试题的难度 p_H；

(3) 从得分最低的一份试卷开始，依次向上选出全部试卷的27%作为低分组，计算出低分组对应的各试题的难度 p_L；

(4) 相应试题的区分度是 p_H 与 p_L 的差，即 $D=p_H-p_L$。

为了与按照相关法计算的区分度加以区别，有人又将极端分组法计算的区分度称为试题的鉴别力指数。这种估计区分度值的方法在日常测验的分析活动中使用较为简便。

三、区分度的评价标准

一般地，试题的区分度是利用相关系数的方法计算得到的，所以，理论上试题区分度的取值在−1.00～1.00之间。如果试题的区分度为负，则说明该试题的功能与整个测验要求相反，当然应该剔除；如果区分度为正，那么区分度到底多高才算好？艾伯尔就选择题区分度值提出了评价标准，具体要求如表 2-7。对于有些要求不高的测验，有些试题的区分度低一些也是容许的。①

---

① 漆书青，现代测量理论在考试中的应用，武汉：华中师范大学出版社，2006，98 页。

表 2-7 试题区分度的评价标准

区分度	评　　价
0.40 以上	优秀
0.30～0.39	良好，如能改进更好
0.20～0.29	尚可，用时需作改进
0.19 以下	劣，必须淘汰或改进以提高区分度后方可使用

在实施完测验后，根据表 2-7 解释试题的区分度时，需要注意以下两类问题：

1. 教育测验中，区分度值一般为正

计算试题区分度时，如果出现区分度值为负数的现象，则暗示该试题可能测量的是不同的心理结构，不适用目前所要测量的心理结构。

2. 当试题区分度小于 0.29 甚至 0.2 时，不一定说明试题不好

造成试题的区分度偏小的原因很多，需要具体问题具体分析。分析时，以下几种情况需要综合考虑。

（1）参加测验人数的多少。如果参加测验人数太少，可能会导致区分度值不稳定。

（2）试题的难度。同一道试题对于不同考生群体可能难度差异很大，如果试题难度过大或过小，都可能导致区分度降低。

（3）试题命制质量。试题的结构、表述等都可能对考生作答造成干扰，影响考生的临场发挥，从而影响试题的区分度。

如果能够排除上述三个方面的因素，那么就需要从教师教学或学生复习等方面来反思教学实践中可能存在的问题。

四、区分度与难度、信度的关系

1. 区分度与难度

每个试题都具有难度、区分度两个技术指标，在计算出测验卷

中所有试题的难度、区分度值后，可以将每个试题的这两个技术指标作为点（难度、区分度），在图2－1所示的难度值为横轴、区分度为纵轴的直角坐标系中标示出来，进行质量性能的综合评价。一般而言，点落在区域Ⅳ以外的题目都要引起注意，尤其是落在区域Ⅰ与区域Ⅲ中的试题必须认真分析出现这种情况的原因，务必保证以后测验中遇到类似的题目必须修改或淘汰。

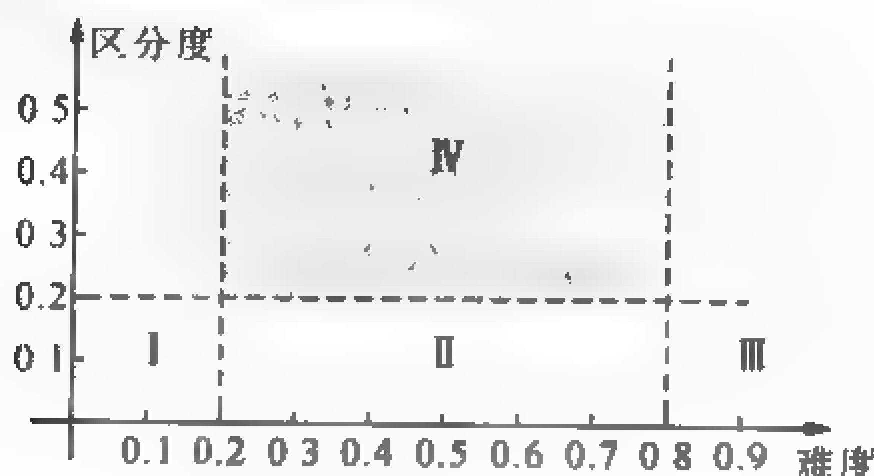


图2－1

另外，对常模参照性测验来说，试题的难度在0.5左右时，区分度较好；难度接近1或0的试题，无区分度可言。对目标参照性测验而言，试卷的区分度意义不大。

表2－8　区分度的最大值与难度的关系

难度(P)	1.00	0.90	0.70	0.50	0.30	0.10
区分度的最大值	0.00	0.20	0.60	1.00	0.60	0.20

由表2－8可知，难度适中的试题区分度最大。

2. 区分度与信度

试题的区分度与测验卷的信度之间也有着紧密的联系。表2－9是艾伯尔1972年发表的，这里试卷信度是在假定全部试题难度均为0.50的前提下预测得到的，表中的区分度指的是试卷中所有试题区分度的平均值。

表 2－9　区分度与测验信度的关系

区分度	信度	区分度	信度
0.1225	0.00	0.30	0.84
0.16	0.42	0.40	0.915
0.20	0.63	0.50	0.949

根据表 2－9 可知，当试题难度确定时，随着试卷区分度增大，测验信度也增大。可见，要想达到理想的测验信度，提高区分度是一个好方法。

五、影响试卷区分度的因素

影响测验卷与试题区分度的因素有很多，如，试题的难度；整卷的难度分布；试题得分点的层次性；试题的解题方法的多样性，等等。

一般而言，若要提高试题的区分度，则尽可能多地考查解答过程较为复杂的问题，使得能力高的考生能得高分，能力差的考生得低分，考生成绩尽量分布在整个分数量尺上。

第五节　EXCEL 与 SPSS 软件应用实例

随着计算机技术的发展与家用电脑的普及，绝大多数教师都会使用微软办公系统（Microsoft Office）中的 EXCEL 软件处理简单的数据与图表，而汉化版的 SPSS 软件含义清晰，操作简便，因此，学会使用 EXCEL 软件与汉化版的 SPSS 软件进行简单的试卷质量分析也很容易。

一、计算测验信度

【例 2－1】　初三A 班共有 10 名学生，表 2－10 是 A 班一次数

表 2-10　A 班一次数学单元测验的原始成绩统计表

考生学号	试题 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	总分
1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	9	9	10	10	12	12	12	14	10	146
2	3	3	3	3	3	3	3	0	0	3	3	3	0	0	3	3	9	9	10	10	1	12	12	11	8	118
3	3	3	3	3	3	3	3	3	3	3	0	3	3	3	3	0	9	9	10	10	12	12	12	14	5	135
4	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	9	9	10	10	12	3	12	14	4	131
5	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	9	9	10	10	1	12	12	14	8	133
6	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	9	9	10	10	12	12	12	14	9	145
7	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	9	9	10	10	12	12	12	12	8	142
8	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	9	9	10	10	12	12	6	14	4	134
9	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	0	9	9	9	10	12	12	12	14	4	136
10	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	9	9	10	10	12	12	11	14	6	141

学期末测验的原始成绩统计表，该数学期末测验卷由 25 道题组成，其中 1～10 题为选择题，每题 3 分；11～16 题为填空题，每题 3 分；17～25 题，为解答题，分值依次为 9、9、10、10、12、12、12、14、14 分；试卷满分为 150 分。试计算该次测验的信度。

解： 由于整份测验卷中包含填空题、选择题、解答题等多种题型，且每题的分值也不尽相同，因此，选择使用公式 2.6 计算 α 系数估计测验信度。

方法 1： 利用 EXCEL 软件计算，分为三步进行。

第一步，计算每道题目的标准差；

标准差的计算采用函数“STDEV”，在单元格 B12 中键入“=STDEV(B2:B11)”，按 Enter 键，返回值 0.00 就显示在单元格 B12 中。B12 表示的是第 1 题实测分数的标准差。

将光标放在单元格 B12 右下角直到显示为“+”，按住左键拖动光标，就可以得出第 2～25 题以及测验总分的标准差。如图 2－2 所示。

B12 =STDEV(B2:B11)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	学号	1	2	3	4	5	6	7	8	[illegible]	10	11	12	13	14	15	16	17
2	1	3	[illegible]	3	3	3	3	3		[illegible]	[illegible]	3	3	[illegible]	3	3	3	9
3	2	3	3	3	3	3	3	3	[illegible]	[illegible]	3	3	5	[illegible]	0	3	3	9
4	[illegible]	3	3	3	3	3	3	3	2	[illegible]	[illegible]	0	[illegible]	3	3	3	[illegible]	9
5	4	3	3	3	3	3	3	3	[illegible]	3	3	3	3	3	[illegible]	3	3	9
6	5	3	3	3	3	3	3	3	3	[illegible]	[illegible]	[illegible]	3	3	3	3	3	9
7	6	3	3	3	3	[illegible]	[illegible]	3	3	[illegible]	3	3	3	3	3	[illegible]	3	9
8	7	3	3	3	3	3	3	3	3	[illegible]	3	3	3	3	3	[illegible]	3	9
9	8	3	3	3	3	3	3	3		3	[illegible]	3	3	3	3	[illegible]	3	9
10	9	3	3	3	[illegible]	3	3	3	3	[illegible]	3	3	3	3	3	3	0	9
11	10	3	3	3	3	[illegible]	3	3	[illegible]		3	[illegible]	[illegible]	3	3	3	3	3
12	标准差	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.45	0.95	0.00	0.95	0.00	0.95	[illegible]95	0.00	1.26	[illegible]

图 2－2

第二步，计算每道题目的方差；

方差等于标准差的平方，在单元格 B13 中键入“=B12 * B12”，按 Enter 键，返回值 0.00 就显示在单元格 B13 中。B13 表示的是第 1 题实测分数的方差，如图 2－3 所示。

将光标放在单元格 B13 右下角直到显示为“+”，按住左键拖动光标，就可以得出 2～25 题以及测验总分的方差。如图 2－3 所示。

图 2－3

第三步，计算测验卷信度的内部一致性系数。

根据公式 2.6，用 $\alpha=\frac{25}{25-1}\left(1-\frac{\sum_{i=1}^{25}\sigma_{x_i}^2}{\sigma_x^2}\right)$ 来估计测验信度。在单元格 B14 中键入“＝(1－SUM(B13：Z13)/AA13)＊25/24”，按 Enter 键，返回值 0.33 就显示在单元格 B14 中。如图 2－4 所示。

图 2－4

所以，该测验信度的内部一致性系数约为 0.33。

方法 2：利用 SPSS 软件计算，分两步进行。

第一步，在 SPSS 数据编辑器中单击“ ”，找到保存为扩展名是“.xls”的电子表格文件，然后根据提示，把在 EXCEL 软件中整理的考生各题的得分数据与总分数据导入 SPSS 数据编辑器中（本书使用的是 SPSS12.0Windows 汉化版），如图 2－5 所示，并将之保存为例 2－1.sav 文件。

图 2-5

第二步，执行【分析】/【尺度分析】/【可靠性分析】程序，如图 2-6。

图 2-6

接着，执行一系列操作。单击"可靠性分析"后，出现"可靠性分析"对话框，如图 2-7，将 x1～x25 选入右边的"项"方框中，"模型"

图 2-7

栏选择"α";然后单击"统计量"按钮,出现"可靠性分析:统计量"子对话框,如图2-8,在"项之间"方框中,选取"相关性",单击"继续"按钮回到"可靠性分析"对话框,单击"确定"按钮,得出结论。

图 2-8

结论中出现"警告:在分析中计算并使用协方差矩阵。以下每一个分变量均有为零的方差,并且已从刻度中移去:x1, x2, x3, x4, x5, x6, x7, x10, x12, x15, x17, x18, x20",并给出下表2-10。

表 2-11 可靠性统计量

Cronbach's Alpha	基于标准化项的 Cronbachs Alpha	项数
0.349	0.646	12

从表2-11可以看出,科隆巴赫α系数值是0.349,去除13道题后,科隆巴赫α系数值是0.646。

【说明】 由于考生样本过小,只有10人,因此用这个小样本的实测分数来估计测验信度显然数值偏低。

【例2-2】 初三年级共有439人,测验卷的结构与例2-1中的相同,求该次期末的数学测验信度。

解法1: 利用EXCEL软件,与例2-1中的解法1同样的步骤

计算，如图 2－9 所示：

B443　=... SUM(B442 Z442)/AA442)*439/428

	学号	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17	x18	x19	x20	x21	x22	x23	x24	x25	总分数
2	1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	9	9	10	10	12	12	12	14	10	146
3	2	3	3	3	3	3	3	3	0	0	3	3	3	0	0	3	3	9	9	10	10	1	12	12	11	8	118
4	3	3	3	3	3	3	3	3	3	[illegible]	3	0	3	3	3	3	0	9	9	10	10	12	12	12	14	5	135
5	4	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	9	9	10	10	12	3	12	14	4	131
6	5	3	3	3	3	3	3	3	5	3	3	3	3	3	3	3	3	9	9	10	10	1	12	12	14	8	133
440	439	0	3	0	3	3	0	0	3	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[illegible]
441	标准差	0.75	0.8	1.2	1.3	0.7	1.2	1.2	1.2	1.5	1.2	1.4	1.4	1.6	1.4	1.4	1.5	3.4	3.6	4.1	4.2	5.5	4.1	5.6	5.9	3	42.88
442	方差	0.56	0.6	1.6	1.7	0.5	1.5	1.4	1.6	2.3	1.5	2	1.9	2.1	2	2	2.3	12	13	17	18	30	17	32	34	10	1839
443	信度	0.91																									

图 2－9

所以，该 2008 学年初三数学期末测验信度的内部一致性系数约为 0.91。

解法 2：利用 SPSS 软件，与例 2－1 中的解法 2 同样的步骤计算，文件保存为“例 2－2.sav”，得到结果如下：

表 2－12　可靠性统计量

Cronbach's Alpha	基于标准化项的 Cronbachs Alpha	项数
0.925	0.942	25

【说明】　对比例 2－1 与例 2－2，当考生群体增大时，测验信度也显然增大；同时，由于考生的水平分布广，也提高了测验信度。

二、计算测验效度

【例 2－3】　初三年级共有 439 人，测验卷的结构与例 2－1、2－2 中的相同，请用 SPSS 软件分析该次期末数学测验的结构效度。

解：第一步，打开例 2－2.sav 文件。

第二步，执行【分析】/【数据降维】/【因子分析】程序，如图 2－10。单击“因子分析”按钮，出现“因子分析”对话框，将左边方框中的题目 x1～x25 选入右边的“变量”下的空框中，如图 2－11。

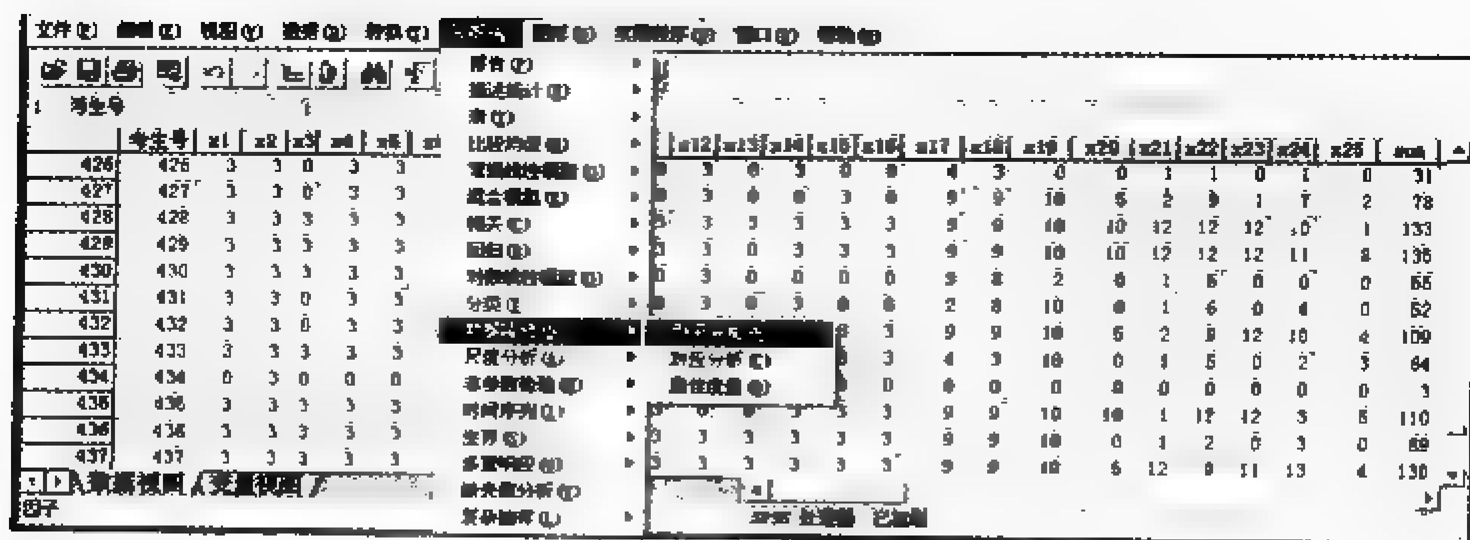


图 2－10

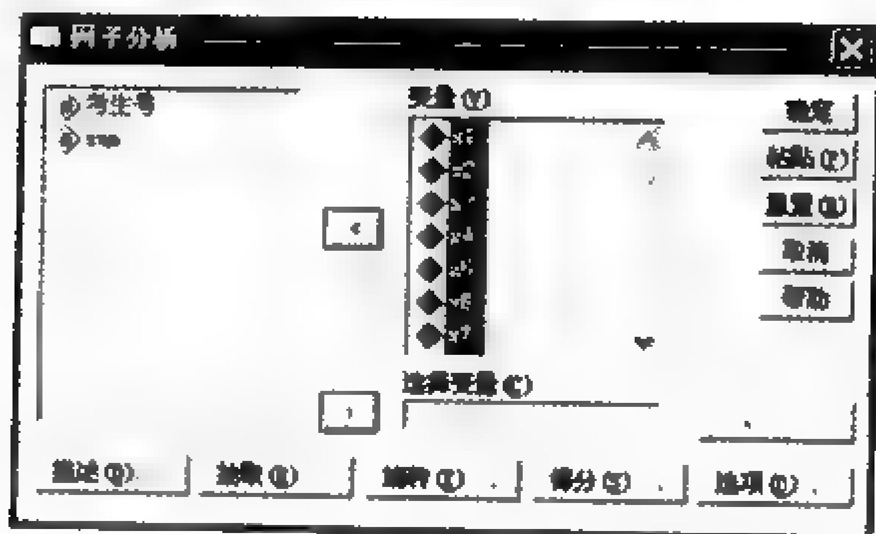


图 2－11

第三步，对“因子分析”中的五个按钮“描述”、“抽取”、“旋转”、“得分”、“选项”中的选项进行选择，基本的选择项与选择方法如图 2－12、2－13、2－14、2－15、2－16 所示，每个选项选择完后，单击“继续”按钮，返回图 2－11。

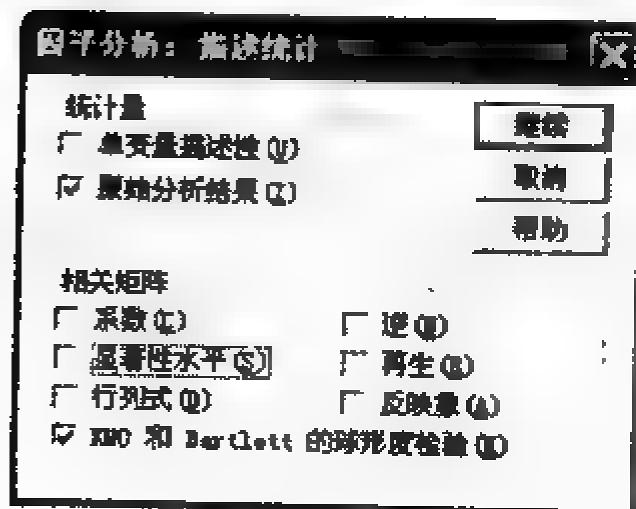


图 2－12

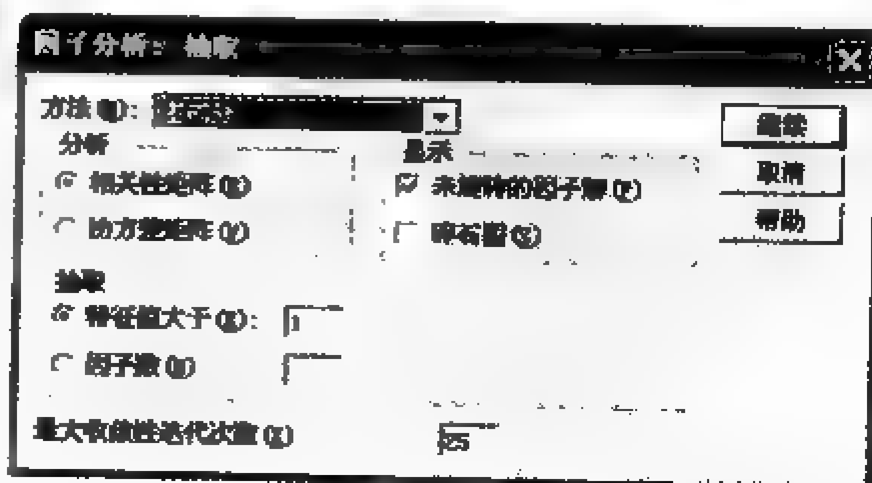


图 2　13

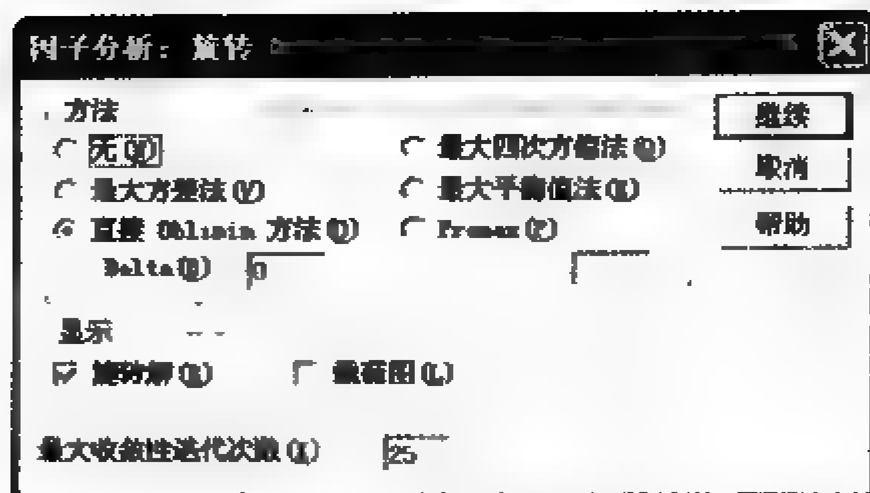


图 2-14



图 2-15

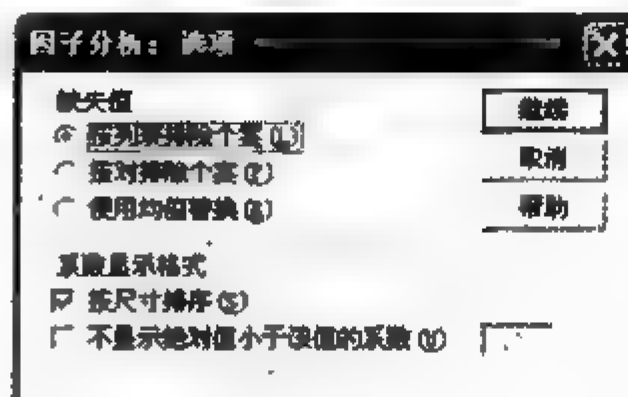


图 2-16

第四步，当“因子分析”的相关选项都完成后，按图 2-11 中的“确定”按钮，就得到一系列的因素分析结果。

表 2-13　KMO 和 Bartlett 的检验

取样足够度的 Kaiser-Meyer-Olkin 度量		0.966
Bartlett 的球形度检验	近似卡方	5872.588
	df	300
	Sig.	0.000

根据表 2-13，KMO 值为 0.966，表明本测验非常适合进行因素分析（判别标准：KMO≥0.9 时，非常适合；0.8＜KMO≤0.9 时，适合；0.7＜KMO≤0.8 时，一般；0.5＜KMO≤0.7 时，不太适合；KMO≤0.5 时，不适合①）；Bartlett 球形度检验的χ^2值为 5872.588

---

① 杨晓明主编，SPSS 在教育统计中的应用，北京：高等教育出版社，2004 年 5 月第一版。

(自由度为300)达到显著水平,表示测验试题中共同因素存在,适合进行因素分析。

表2 14 公因子方差(提取方法:主成分分析)

	初始	提取		初始	提取		初始	提取
x1	1.000	.241	x10	1.000	.267	x18	1.000	.742
x2	1.000	.435	x11	1.000	.580	x19	1.000	.680
x3	1.000	.419	x12	1.000	.291	x20	1.000	.760
x4	1.000	.558	x13	1.000	.608	x21	1.000	.643
x5	1.000	.400	x14	1.000	.374	x22	1.000	.719
x6	1.000	.508	x15	1.000	.477	x23	1.000	.716
x7	1.000	.527	x16	1.000	.418	x24	1.000	.792
x8	1.000	.522	x17	1.000	.721	x25	1.000	.723
x9	1.000	.450						

表2-14中,利用主成分分析法,得到25个题项的共同性。共同性越高,表示该题与其他题目可测量的共同特质越多,即该题的重要性越大。显然解答题x17～x25的重要性比选择题x1～x10、填空题x11～x16都高。

表2-15 解释的总方差

成分	初始特征值			提取平方和载入			旋转平方和载入(a)
	合计	方差的%	累积%	合计	方差的%	累积%	合计
1	10.850	43.400	43.400	10.850	43.400	43.400	9.529
2	1.636	6.546	49.946	1.636	6.546	49.946	5.839
3	1.087	4.348	54.294	1.087	4.348	54.294	5.386
4	.940	3.760	58.054				
5	.864	3.456	61.510				
6	.844	3.375	64.885				

（续表）

成分	初始特征值			提取平方和载入			旋转平方和载入(a)
	合计	方差的%	累积%	合计	方差的%	累积%	合计
7	.798	3.191	68.075				
8	.746	2.984	71.060				
9	.712	2.848	73.908				
10	.678	2.714	76.621				
11	.632	2.526	79.147				
12	.567	2.268	81.416				
13	.539	2.156	83.571				
14	.502	2.008	85.580				
15	.461	1.844	87.423				
16	.430	1.720	89.144				
17	.417	1.667	90.810				
18	.397	1.587	92.397				
19	.361	1.443	93.840				
20	.320	1.279	95.118				
21	.307	1.229	96.348				
22	.267	1.069	97.417				
23	.239	.957	98.374				
24	.209	.837	99.211				
25	.197	.789	100.000				

提取方法：主成分分析。

a 使成分相关联后，便无法通过添加平方和载入来获得总方差。

表2-15中，初始特征值栏目中，“合计”栏为特征值，共有25个因素，与试题总数相等，表示的是每个成分解释的试题方差总量；“方差的%”为解释的方差比例，例如成分1解释的方差占总方差的43.4%，如果因子分析效果好，应该出现很少数的因子解释很大比例的方差的情况；“累积%”为解释的累积方差比例，可见前三个成分解释的方差占了总方差的54.294%。“提取平方和载入”栏目中，

以特征值等于1为抽取标准，共抽取了3个主成分(公共因子)。“旋转平方和载入”栏目中，由于执行的是斜交旋转，公共因子间是相关的，具有公共方差，因此，解释的方差无法在公共因子之间做分配，但从“合计”一栏可以看出，旋转后，三个主成分之间特征值的差异明显缩小，由原来的“10.850、1.636、1.087”变为“9.529、5.839、5.386”。

表2-16是保留了三个主成分后的因子负荷矩阵，由于主成分1的可解释的方差占总方差的比例高达43.4%，而主成分2、主成分3的可解释的方差占总方差的比例却分别仅为6.546%、4.348%，与主成分1相差过大，因此需要对主成分进行旋转后重新提取。执行斜交旋转后，得到模式矩阵如表2-17，它显示的是旋转后，每道试题在三个主成分上的负荷，每个负荷值实际上就是相应的试题与旋转后的因子的偏相关系数。

转轴后的三个公共因子，每个因子的题目数比较适当，三个因子所包含的题项分别为：

因子1：第9、10、11、12、13、14、15、16、20、21、23、24、25题；

因子2：第1、2、5、17、18、19、22题；

因子3：第3、4、6、7、8题。

结合测验卷中具体的试题，可以给因子1命名为“综合推理的方法与技巧”，因子2命名为“基本概念与法则的理解与运用”，因子3命名为“基本计算技巧的使用”，那么该份测验卷考查考生数学能力主要从这三个方面进行的。

表2-18与表2-17的形式相同。如果执行的是正交旋转，则模式矩阵与结构矩阵相等；如果执行的是斜交旋转，结构矩阵比模式矩阵复杂，这时，矩阵中的相关系数既受试题与成分间的相关关系影响，又受到因子间的相关关系影响，具体数据的计算可以结合表2-17与表2-19共同进行，计算方法可以参考相关的SPSS使用指导书。

表 2-16 成分矩阵(a)

	主成分(公共因子)				主成分(公共因子)				主成分(公共因子)		
	1	2	3		1	2	3		1	2	3
x20	.849	−.132	−.144	x22	.714	.321	−.326	x9	.550	−.320	.211
x24	.848	−.235	−.134	x15	.688	−.055	.021	x8	.540	.038	.479
x23	.809	−.228	−.098	x13	.646	−.436	.026	x12	.517	−.127	.086
x18	.793	.280	−.188	x6	.639	.157	.273	x10	.510	−.016	.080
x17	.788	.229	−.219	x16	.631	−.118	.079	x2	.488	.418	.147
x11	.761	.013	−.007	x7	.627	.316	.187	x3	.488	.226	.360
x25	.760	−.368	−.097	x4	.624	−.019	.411	x1	.349	.343	−.038
x19	.745	.180	−.304	x14	.593	−.116	−.097	x5	.438	.453	−.057
x21	.735	−.308	−.090								

提取因子方法:主成分分析法。

a 已提取了 3 个主成分。

表 2-17　模式矩阵(a)

	主成分(公共因子)				主成分(公共因子)				主成分(公共因子)		
	1	2	3		1	2	3		1	2	3
x25	.889	−.047	−.047	x22	.347	.702	.127	x8	.149	.050	.659
x24	.849	.125	−.034	x18	.382	.615	.033	x4	.284	−.039	.585
x13	.820	−.213	.048	x5	−.069	.592	.151	x3	−.004	.172	.566
x21	.817	−.003	−.029	x17	.436	.582	−.017	x6	.197	.204	.482
x23	.802	.099	.000	x19	.483	.569	−.133	x7	.084	.393	.426
x20	.764	.226	−.017	x2	−.086	.466	.380				
x9	.577	−.235	.265	x1	−.042	.451	.124				
x14	.553	.133	−.014								
x11	.520	.260	.156								
x15	.518	.157	.154								
x16	.510	.049	.190								
x12	.436	.001	.171								
x10	.336	.106	.192								

提取方法：主成分分析法。
旋转法：具有 Kaiser 标准化的斜交旋转法。
a　旋转在 16 次迭代后收敛。

表 2-18 结构矩阵

	主成分(公共因子)				主成分(公共因子)				主成分(公共因子)		
	1	2	3		1	2	3		1	2	3
x24	.883	.457	.420	x22	.571	.796	.298	x8	.445	.253	.712
x25	.848	.297	.362	x18	.647	.782	.442	x4	.549	.291	.707
x20	.847	.530	.433	x17	.664	.753	.407	x6	.511	.461	.651
x23	.842	.424	.421	x19	.649	.716	.308	x3	.338	.379	.628
x21	.802	.318	.362	x5	.243	.619	.335	x7	.448	.584	.611
x13	.756	.137	.363	x2	.286	.571	.510				
x11	.701	.528	.501	x1	.200	.480	.270				
x15	.656	.424	.460								
x16	.621	.325	.452								
x9	.609	.097	.456								
x14	.600	.352	.300								
x12	.518	.240	.380								
x10	.471	.313	.393								

提取方法：主成分分析法。
旋转法：具有 Kaiser 标准化的斜交旋转法。

表 2 19 成分转换矩阵

成分	1	2	3
1	1.000	.406	.480
2	.406	1.000	.368
3	.480	.368	1.000

提取方法：主成分分析法。
旋转法：具有 Kaiser 标准化的斜交旋转法。

在使用斜交旋转时，由于结构矩阵的解释非常复杂，一般情况下，只是检查模型矩阵的因子负荷，并对其作出解释。

表 2－19 为主成分与主成分之间的相关系数矩阵，可以看出主成分之间的相关性比较适当。表 2－20 为成分得分系数矩阵，根据该矩阵，可以计算每个考生在各个主成分上的得分。例如，计算某个考生在成分 1 上的得分，计算方法为：成分 1 的得分 $= \sum_{i=1}^{25} r_{1i} \times x_i$，其中 r_{1i}表示成分 1 与第 i 题的相关系数，在表 2－19 的矩阵中可以查得，x_i 为试题 i 的得分。当测试具有足够的代表性时，这一结果比原始成绩更能具体地反映考生在能力结构上的差异。

三、计算试题的难度

【例 2－4】 初三年级共有 439 人，测验卷与例 2－2、2－3 中的相同，请利用 EXCEL 软件计算全年级该次期末数学测验的每道试题的平均分和难度。

解：在具体操作时，客观题与主观题的难度都采用公式 2.13 来计算。

第一步，在全体考生成绩下面第 441 行放入各题的满分值，如图 2－17 中第 441 行所示。

第二步，计算各题的平均分。平均分的计算采用函数“AVERAGE”，在单元格 B442 中键入“ AVERAGE(B2:B440)”，按 Enter 键，返回值 2.8 就显示在单元格 B442 中。B442 表示的是全年级第 1 题的平均分。

表 2 20 成分得分系数矩阵

	主成分(公共因子)				主成分(公共因子)				主成分(公共因子)		
	1	2	3		1	2	3		1	2	3
x1	−.044	.173	.036	x10	.041	.004	.082	x18	.019	.215	−.044
x2	−.064	.163	.177	x11	.063	.055	.043	x19	.047	.202	−.136
x3	−.038	.027	.296	x12	.066	−.043	.074	x20	.114	.037	−.059
x4	.025	−.079	.310	x13	.151	−.146	.006	x21	.139	−.057	−.051
x5	−.060	.229	.042	x14	.084	.015	−.042	x22	.015	.264	−.136
x6	−.002	.032	.239	x15	.069	.015	.050	x23	.128	−.017	−.043
x7	−.032	.119	.200	x16	.074	−.031	.078	x24	.136	−.008	−.065
x8	.000	−.079	.357	x17	.033	.202	−.071	x25	.155	−.078	−.062
x9	.102	−.154	.137								

提取方法:主成分分析法。
旋转法:具有 Kaiser 标准化的斜交旋转法。

将光标放在单元格 B442 右下角直到显示为“+”，按住左键拖动光标，就可以得出全年级第 2～25 题以及测验卷的平均分。如图 2－17 中第 442 行所示。

B442 fx AVERAGE(B2:B440)

	A	B	C	D	E	F	G	H	I	J	S	T	U	V	W	X	Y	Z	AA
1	考生号	x1	x2	x3	x4	x5	x6	x7	x8	x9	x18	x19	x20	x21	x22	x23	x24	x25	sum
2	1	3	3	3	3	3	3	3	3	3	9	10	10	12	12	12	14	[illegible]	146
3	2	3	3	3	3	3	3	3	0	0	9	10	10	1	12	12	11	8	118
439	438	3	3	0	0	3	3	0	0	0	0	2	0	0	12	0	0	0	29
440	439	0	3	0	3	3	0	0	3	0	0	0	0	0	0	0	0	[illegible]	15
441	满分	3	3	3	3	3	3	3	3	3	9	10	10	12	12	12	14	14	150
442	平均分	2.8	2.77	2.36	2.22	2.82	2.39	2.4	2.34	1.54	6.87	7.44	5.77	5.18	9.42	6.29	7.22	2.81	92.51
443	难度	0.93	0.92	0.79	0.74	0.94	0.8	0.8	0.78	0.51	0.76	0.74	0.58	0.43	0.79	0.52	0.52	0.2	0.617
444																			

图 2－17

第三步，计算各题的难度。在单元格 B443 中键入“=B442/B441)”，按 Enter 键，返回值 0.93 就显示在单元格 B443 中。B443 表示的是全年级第 1 题的难度。

将光标放在单元格 B443 右下角直到显示为“+”，按住左键拖动光标，就可以得出全年级第 2～25 题以及测验卷的难度。如图 2－26 中第 443 行所示。

四、计算试题的区分度

【例 2－5】 初三年级共有 439 人，测验卷与例 2－2、2－3、2－4 中的相同，请利用 EXCEL 软件计算全年级该次期末数学测验的每道试题的区分度。

解：本题的区分度计算采用高、低分组的方法，用 EXCEL 软件分四步进行。

第一步，确定全年级 439 人中高分组与低分组的人数，即 439×27%=180 人。

第二步，将全年级考生按照测验总分由高到低排序。

如图 2－18，执行【数据】/【排序】程序，出现图 2－19 的“排序”

对话框。在“主要关键字”栏目选择“sum”与“降序”，即首先按照总分把数据由高到低降序排列。单击“确定”按钮，即得到排序后的考生数据。

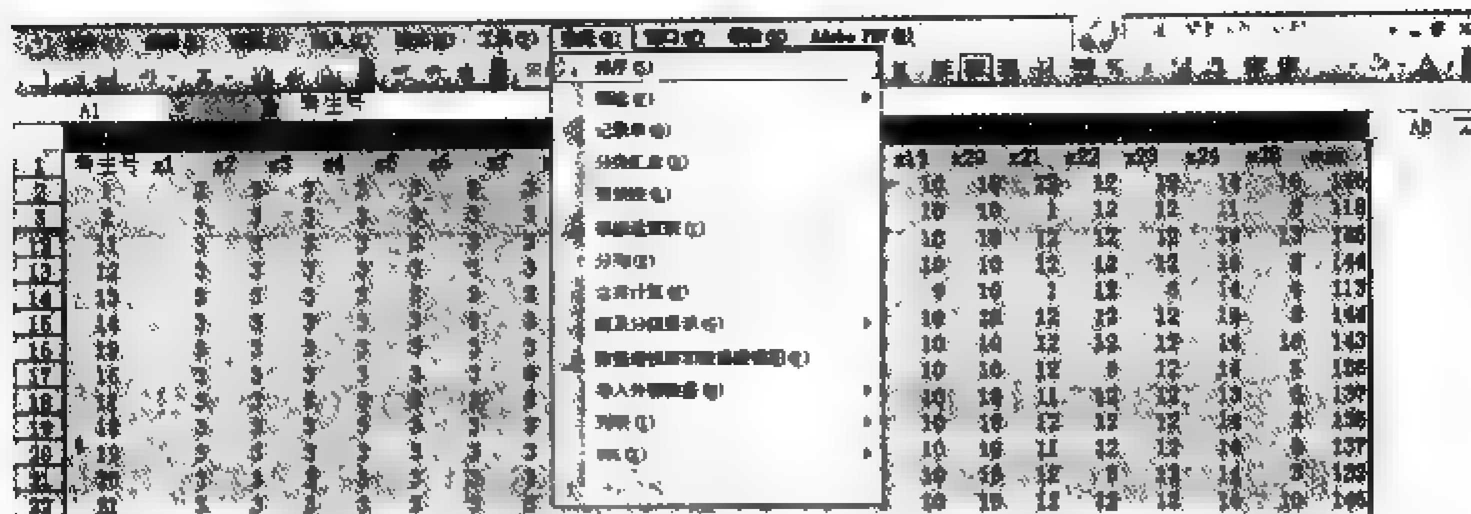


图 2－18



图 2－19

第三步，计算各题和总分的高分组难度与低分组难度。

在单元格 B442 中键入“=SUM(B3:B182)/(180 * B2)”，按 Enter 键，返回值 1.00 就显示在单元格 B442 中。B442 表示的是高分组 180 名考生第 1 题的难度。

将光标放在单元格 B442 右下角直到显示为“十”，按住左键拖动光标，就可以得出高分组 180 名考生第 2～25 题以及测验总分的难度。如图 2－20 所示。

B442	=SUM(B3:B182)/(180*B2)

	A	B	C	D	E	F	G	H	I	J	P	Q	S	T	U	V	W	X	Y	Z	AA	AB
1	考生号	x1	x2	x3	x4	x5	x6	x7	x8	x9	x15	x16	x18	x19	x20	x21	x22	x23	x24	x25	sum	名次
2	各题满分	3	3	3	3	3	3	3	3	3	3	3	9	10	10	12	12	12	14	14	150	
3	92	3	3	3	3	3	3	3	3	3	3	3	9	10	10	12	12	12	14	12	148	1
4	118	3	3	3	3	3	3	3	3	3	3	3	9	10	10	12	12	12	14	12	148	2
13	11	3	3	3	3	3	3	3	3	3	3	3	9	10	10	12	12	12	14	13	146	11
437	320	3	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	6	435
438	323	0	0	0	3	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	6	436
439	392	0	3	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	6	437
440	434	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	438
441	296	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	439
442	H180难度	1	1	0.93	0.98	0.99	0.99	0.97	0.96	0.92	0.96	0.21	0.99	0.98	0.93	0.87	0.97	0.96	0.9	0.4	0.891	
443	L180难度																					

图 2－20

在单元格 B443 中键入“＝SUM(B260:B439)/(180 * B2)”，按 Enter 键，返回值 0.9 就显示在单元格 B443 中。B443 表示的是低分组 180 名考生第 1 题的难度。

将光标放在单元格 B443 右下角直到显示为“＋”，按住左键拖动光标，就可以得出低分组 180 名考生第 2～25 题以及测验总分的难度。如图 2－21 所示。

B443	=SUM(B260:B439)/(180*B2)

	A	B	C	D	E	F	G	H	I	J	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1	考生号	x1	x2	x3	x4	x5	x6	x7	x8	x9	x15	x16	x17	x18	x19	x20	x21	x22	x23	x24	x25	sum	名次
2	各题满分	3	3	3	3	3	3	3	3	3	3	3	9	9	10	10	12	12	12	14	14	150	
3	92	3	3	3	3	3	3	3	3	3	3	3	9	9	10	10	12	12	12	14	12	148	1
4	118	3	3	3	3	3	3	3	3	3	3	3	9	9	10	10	12	12	12	14	12	148	2
13	11	[illegible]	3	3	3	3	3	3	3	3	3	3	9	9	10	10	12	12	12	14	13	146	11
438	323	0	0	0	[illegible]	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	6	436
439	392	0	3	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	437
440	434	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	438
441	296	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	439
442	H180难度	1	1	0.93	0.98	0.99	0.99	0.97	0.98	0.82	0.96	0.83	0.98	0.99	0.98	0.93	0.87	0.97	0.96	0.9	0.4	0.89	
443	L180难度	0.9	0.82	0.63	0.46	0.87	0.54	0.59	0.58	0.22	0.34	0.17	0.45	0.46	0.43	0.19	0.06	0.56	0.07	0.11	0.02	0.32	
444																							

图 2－21

第四步，计算各题与测验卷的区分度。

在单元格 B444 中键入“　B442－B443”，按 Enter 键，返回值 0.1 就显示在单元格 B444 中。B444 表示的是第 1 题的区分度。

将光标放在单元格 B444 右下角直到显示为“＋”，按住左键拖动光标，就可以得出第 2～25 题以及测验总分的区分度。如图2　22 所示。

图 2－22

【例 2－6】 初三年级共有 439 人，测验卷与例 2－2、2－3、2－4、2－5 中的相同，请利用 SPSS 软件用高低分组法分析全年级该次期末数学测验每道试题的区分度，并检验其显著性。

解：区分度的分析分为四步进行。

第一步，确定全年级 439 人中高分组与低分组的人数，即 439×27％＝180 人，然后确定高分组的分数分布是 116 分至最高分，低分组的分数分布是最低分至 87 分。

第二步，将全年级考生按照测验总分由高到低排序。

如图 2－23，执行【数据】/【对个案排序】程序，出现图 2－24 的“个案排序”对话框。将左边变量名“sum”选中放入右边“排序方式”下的方框中，在“排序顺序”中选择“降序”方式，单击“确定”按钮，即把全部数据按照总分进行了由高到低的排列。

图 2－23

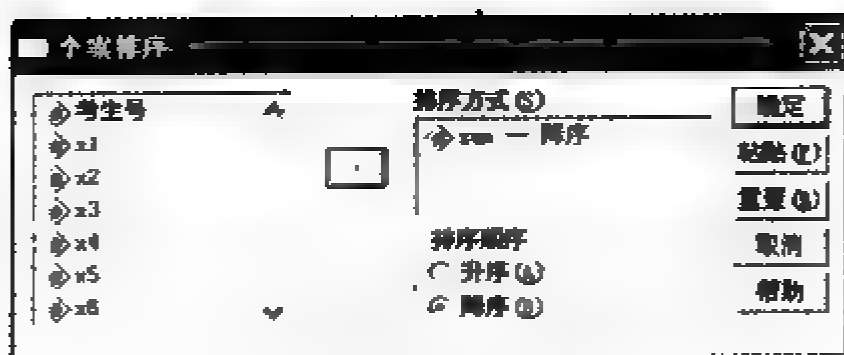


图 2-24

第三步，选出高、低分组 27%的分数，作为高、低分组的界限(高分组为第 1 组，低分组为第 2 组)。

如图 2-25，执行【转换】/【重新编码】/【成不同变量】程序，出现图 2-26 的"重新编码为其他变量"对话框。将左边变量名"sum"选中放入右边"数字变量→输出变量"下的方框中，在"输出变量"栏目"名称"中输入"group"，单击"更改"按钮，"数字变量→输出变量"下方框中的内容由原来的"sum→?"变为"sum→group"。

图 2-25

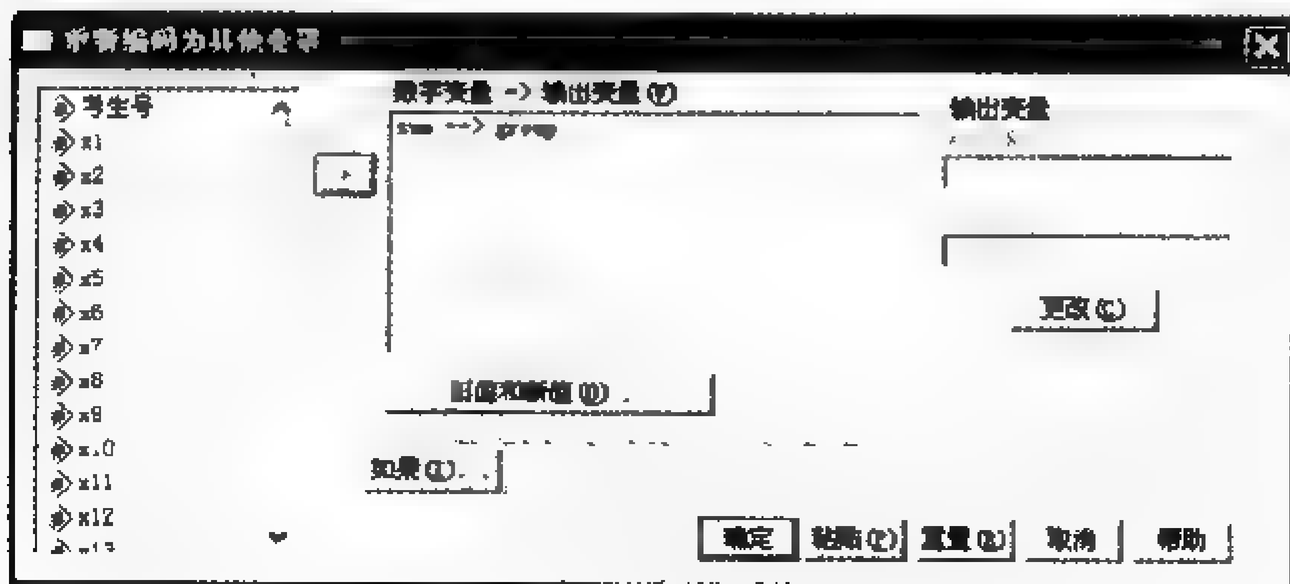


图 2-26

单击“旧值和新值(Q)…”按钮，出现“重新编码到其他变量：旧值和新值”的子对话框，如图2-27所示。

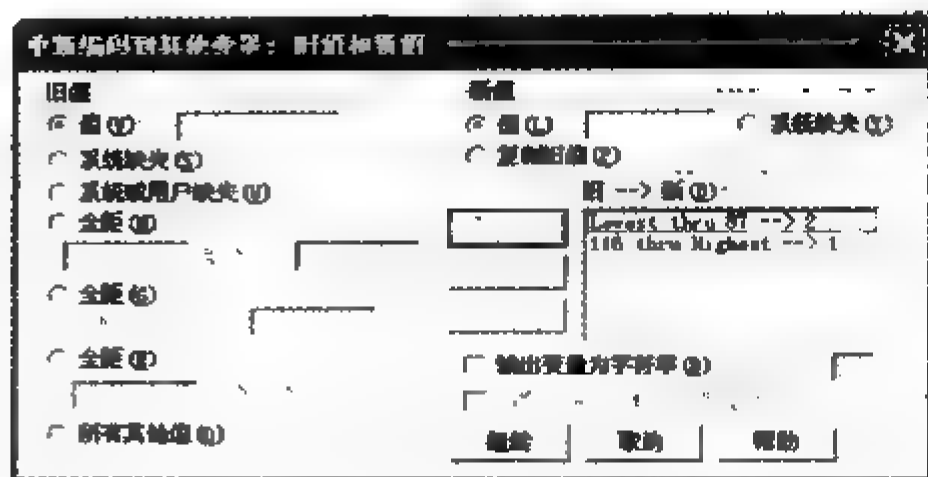


图 2-27

在左边“旧值”方框中，由上到下选择第5个选项“全距(G)：”，在“从最小值到”后面的空格内输入低分组界限“87”(表示最低分至87分)，在右边“新值”方框中，选择“值”，在后面的空格内输入“2”(低分组设为第2组)，然后单击“添加”按钮，在“旧→新”下的方框中出现“Lowest thru 87→2”，表示数据中总分在87分以下的考生为低分组(第2组)。

同样地，在左边“旧值”方框中，由上到下选择第6个选项“全距(G)：”，在“到最大值”前面的空格内输入高分组界限“116”(表示116分至最高分)，在右边“新值”方框中，选择“值”，在后面的空格内输入“1”(高分组设为第1组)，然后单击“添加”按钮，在“旧→新”下的方框中出现“116 thru 87 Highest→1”，表示数据中总分在116分以上的考生为高分组(第1组)。如图2-27所示。

图2-27中，单击“继续”按钮，回到“重新编码到其他变量：旧值和新值”的子对话框，再单击“确定”按钮，数据文件的窗口中新增一个“group”的变量，变量值为1(高分组)或2(低分组)。

第四步，用T检验来检验测验卷高、低两组在各个试题上的差异。

如图2-28，执行【分析】/【比较均值】/【独立样本T检验】程序，出现图2-29的“独立样本T检验”对话框。将左边变量“x1～x25，sum”选中放入右边“检验变量”下的方框中。

图 2－28

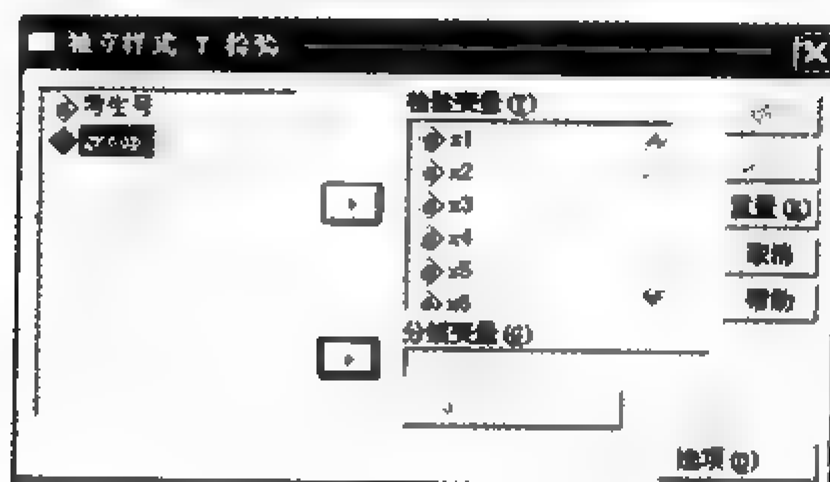


图 2－29

将图 2－29 左边变量“group”选中放入右边“分组变量”下的方框中，出现图 2－30 的“group(??)”，单击“定义组”按钮，出现“定义组”子对话框”，选择“使用指定值”，在“组 1”后的空格中输入 1，在“组 2”后的空格中输入 2，单击“继续”按钮，回到“独立样式 T 检验”对话框，单击“确定”按钮。



图 2－30

在“输出－SPSS 浏览器”中出现以高低两组为自变量、以 x1～x25 为因变量所进行的独立样本 T 检验的结果，如表 2－21 与表 2－22 所示。

表 2－21　组统计量

题号	高低分组	人数	均值	标准差	均值的标准误
x1	1.00	182	2.97	.314	.023
	2.00	182	2.57	1.053	.078
x2	1.00	182	3.00	.000	.000
	2.00	182	2.44	1.173	.087
x3	1.00	182	2.80	.747	.055
	2.00	182	1.86	1.460	.108
x4	1.00	182	2.92	.492	.036
	2.00	182	1.37	1.498	.111
x5	1.00	182	2.98	.222	.016
	2.00	182	2.57	1.053	.078
x6	1.00	182	2.98	.222	.016
	2.00	182	1.62	1.500	.111
x7	1.00	182	2.90	.537	.040
	2.00	182	1.75	1.484	.110
x8	1.00	182	2.92	.492	.036
	2.00	182	1.73	1.486	.110
x9	1.00	182	2.46	1.159	.086
	2.00	182	.66	1.246	.092
x10	1.00	182	2.92	.492	.036
	2.00	182	1.78	1.478	.110
x11	1.00	182	2.92	.492	.036
	2.00	182	.91	1.381	.102
x12	1.00	182	2.80	.747	.055
	2.00	182	1.40	1.501	.111

（续表）

题号	高低分组	人数	均值	标准差	均值的标准误
x13	1.00	182	2.27	1.288	.095
	2.00	182	.13	.617	.046
x14	1.00	182	2.72	.875	.065
	2.00	182	1.15	1.464	.108
x15	1.00	182	2.88	.579	.043
	2.00	182	1.02	1.426	.106
x16	1.00	182	2.42	1.186	.088
	2.00	182	.51	1.131	.084
x17	1.00	182	8.84	.818	.061
	2.00	182	4.04	3.714	.275
x18	1.00	182	8.91	.734	.054
	2.00	182	4.09	4.030	.299
x19	1.00	182	9.79	1.053	.078
	2.00	182	4.29	4.555	.338
x20	1.00	182	9.35	1.697	.126
	2.00	182	1.86	2.689	.199
x21	1.00	182	10.44	3.744	.278
	2.00	182	.76	1.547	.115
x22	1.00	182	11.61	1.357	.101
	2.00	182	6.64	4.875	.361
x23	1.00	182	11.38	2.101	.156
	2.00	182	.85	2.500	.185
x24	1.00	182	12.64	2.332	.173
	2.00	182	1.48	2.600	.193
x25	1.00	182	5.61	2.663	.197
	2.00	182	.25	.735	.055
sum	1.00	182	133.42	8.575	.636
	2.00	182	47.73	24.830	1.840

表 2-22 独立样本检验

		方差方程的 Levene 检验		均值方程的 t 检验						
		F 检验	显著性	T 检验	自由度	显著性（双侧）	均值差值	标准误差值	差分的 95% 置信区间	
									下限	上限
x1	假设方差相等	123.700	.000	4.859	362	.000	.396	.081	.235	.556
	假设方差不相等			4.859	212.879	.000	.396	.081	.235	.556
x2	假设方差相等	280.330	.000	6.448	362	.000	.560	.087	.390	.731
	假设方差不相等			6.448	181.000	.000	.560	.087	.389	.732
x3	假设方差相等	363.451	.000	7.732	362	.000	.940	.122	.701	1.179
	假设方差不相等			7.732	269.645	.000	.940	.122	.700	1.179
x4	假设方差相等	1376.009	.000	13.256	362	.000	1.549	.117	1.320	1.779
	假设方差不相等			13.256	219.542	.000	1.549	.117	1.319	1.780
x5	假设方差相等	146.099	.000	5.167	362	.000	.412	.080	.255	.569
	假设方差不相等			5.167	197.122	.000	.412	.080	255	.569
x6	假设方差相等	6275.671	.000	12.174	362	.000	1.368	.112	1.147	1.589
	假设方差不相等			12.174	188.956	.000	1.368	.112	1.146	1.590

（续表）

		方差方程的 Levene 检验		均值方程的 t 检验						
									差分的 95% 置信区间	
		F 检验	显著性	T 检验	自由度	显著性（双侧）	均值差值	标准误差值	下限	上限
x7	假设方差相等	939.294	.000	9.866	362	.000	1.154	.117	.924	1.384
	假设方差不相等			9.866	227.650	.000	1.154	.117	.923	1.384
x8	假设方差相等	1154.108	.000	10.228	362	.000	1.187	.116	.959	1.415
	假设方差不相等			10.228	220.156	.000	1.187	.116	.958	1.416
x9	假设方差相等	3.365	.067	14.245	362	.000	1.797	.126	1.549	2.045
	假设方差不相等			14.245	360.133	.000	1.797	.126	1.549	2.045
x10	假设方差相等	1032.418	.000	9.853	362	.000	1.137	.115	.910	1.364
	假设方差不相等			9.853	220.601	.000	1.137	.115	.910	1.365
x11	假设方差相等	431.786	.000	18.502	362	.000	2.011	.109	1.797	2.225
	假设方差不相等			18.502	226.141	.000	2.011	.109	1.797	2.225
x12	假设方差相等	534.896	.000	11.276	362	.000	1.401	.124	1.157	1.645
	假设方差不相等			11.276	265.400	.000	1.401	.124	1.156	1.646

（续表）

		方差方程的 Levene 检验		均值方程的 t 检验						
		F 检验	显著性	T 检验	自由度	显著性（双侧）	均值差值	标准误差值	差分的 95%置信区间	
									下限	上限
x13	假设方差相等	172.343	.000	20.244	362	.000	2.143	.106	1.935	2.351
	假设方差不相等			20.244	259.844	.000	2.143	.106	1.934	2.351
x14	假设方差相等	243.844	.000	12.388	362	.000	1.566	.126	1.317	1.815
	假设方差不相等			12.388	295.819	.000	1.566	.126	1.317	1.815
x15	假设方差相等	469.198	.000	16.332	362	.000	1.863	.114	1.638	2.087
	假设方差不相等			16.332	239.030	.000	1.863	.114	1.638	2.087
x16	假设方差相等	1.181	.278	15.744	362	.000	1.912	.121	1.673	2.151
	假设方差不相等			15.744	361.195	.000	1.912	.121	1.673	2.151
x17	假设方差相等	404.472	.000	17.016	362	.000	4.797	.282	4.242	5.351
	假设方差不相等			17.016	198.511	.000	4.797	.282	4.241	5.353
x18	假设方差相等	1245.821	.000	15.853	362	.000	4.813	.304	4.216	5.410
	假设方差不相等			15.853	192.984	.000	4.813	.304	4.214	5.412

（续表）

		方差方程的 Levene 检验		均值方程的 t 检验						
		F 检验	显著性	T 检验	自由度	显著性（双侧）	均值差值	标准误差值	差分的 95% 置信区间 下限	上限
x19	假设方差相等	1169.214	.000	15.855	362	.000	5.495	.347	4.813	6.176
	假设方差不相等			15.855	200.285	.000	5.495	.347	4.811	6.178
x20	假设方差相等	49.128	.000	31.799	362	.000	7.495	.236	7.031	7.958
	假设方差不相等			31.799	305.513	.000	7.495	.236	7.031	7.958
x21	假设方差相等	66.445	.000	32.242	362	.000	9.681	.300	9.091	10.272
	假设方差不相等			32.242	241.086	.000	9.681	.300	9.090	10.273
x22	假设方差相等	540.049	.000	13.242	362	.000	4.967	.375	4.229	5.705
	假设方差不相等			13.242	208.880	.000	4.967	.375	4.228	5.707
x23	假设方差相等	3.105	.079	43.510	362	.000	10.533	.242	10.057	11.009
	假设方差不相等			43.510	351.550	.000	10.533	.242	10.057	11.009
x24	假设方差相等	1.137	.287	43.087	362	.000	11.154	.259	10.645	11.663
	假设方差不相等			43.087	357.793	.000	11.154	.259	10.645	11.663

（续表）

		方差方程的 Levene 检验		均值方程的 t 检验						
		F 检验	显著性	T 检验	自由度	显著性（双侧）	均值差值	标准误差值	差分的 95%置信区间	
									下限	上限
x25	假设方差相等	191.693	.000	26.190	362	.000	5.363	.205	4.960	5.765
	假设方差不相等			26.190	208.449	.000	5.363	.205	4.959	5.766
sum	假设方差相等	190.985	.000	44.009	362	.000	85.692	1.947	81.863	89.521
	假设方差不相等			44.009	223.574	.000	85.692	1.947	81.855	89.529

在结果分析中，表 2－20 为高、低两组的描述性统计量，包括组别、个数、平均数、标准差、平均数的标准误，表 2－21 为 T 检验的结果，在查阅报表时，先看每个题组别总体方差相等的“F 检验”，如果 F 值显著（显著性一栏的值小于 0.05），表中假设方差相等不成立，再看“不假设方差相等”栏目的 T 值，如果显著（显著性一栏的值小于 0.05），则表明此题具有鉴别度。显然，上述试题的鉴别度都是极其显著的。

如果 F 值不显著（显著性一栏的值大于 0.05），表明两组的方差总体相等，再看“假设方差相等”栏目的 T 值，如果显著（显著性一栏的值小于 0.05），则表明此题具有鉴别度。

另外，判断两组平均数差异检验的 T 值是否显著，也可以考虑差异值的 95％的置信区间，如果 95％的置信区间为不包括 0 在内，表明两组的差异显著；如果包括 0 在内，表明两组的平均数有可能相等，两者的差异就不显著。

第三章 测验成绩的统计处理

运用教育统计的方法对测验成绩进行处理与分析主要包括两种类别，第一种是描述统计，即把测验成绩及其相关信息进行整理、概括，目的在于将大量零散、杂乱无序的数字资料进行归纳、总结，使事物的全貌及其分布特征清晰、明确地显现出来。第二种是推断统计，即根据测验提供的信息，运用概率的理论进行分析与论证，在一定可靠度上对考生总体能力分布特征进行估计与推测，目的在于根据已知情况推断未知，为下一步决策做准备。这两个类别相互联系，其中描述统计是推断统计的基础，推断统计通过样本信息估计、推测总体，从已知情况估计、推测未知情况。本章主要介绍运用描述统计组织和整理测验成绩，第四章主要介绍运用推断统计分析测验成绩。

第一节　数据的特点与种类

教育测验的结果以数据的形式呈现。由于不同特点和种类的数据，需要采用不同的数据处理和转换的方法，因此，对数据进行处理前，必须首先明确数据的特点和种类。

一、数据的常见三种分类

根据不同的标准可以将数据分为不同的种类。

1. 按照数据的来源分类

按照数据的来源，可以将数据分为点计数据和度量数据。

点计数据是指计算个数所获得的数据。如学校数、班级数、学生数、教师数、考室数、教学仪器数等。

度量数据是指用一定的工具或一定的标准测量所获得的数据。例如，用体重秤测得学生的体重的数据，用时钟测得学生完成某项作业所用时间的数据，用单元测验获得学生该单元知识技能掌握情况的数据，等等。

2. 按照数据的取值特征分类

按照数据的取值特征，可以将数据分为离散型数据和连续型数据。

离散型数据的取值是间断的，数据单位独立，两个单位之间不能再划分成细小的单位，数据一般用整数表示。例如，参加单元测验的男生人数、女生人数；测验获得优、良、中、差各个等级的人数，等等。

连续型数据的取值可以看成来自于实数集合，它们可能的取值范围能连续充满实数集的某一个区间，数据的单位之间可以再划分成无限多个细小的单位，即数据可以用小数表示。例如，学生的身高、体重、智商、测验成绩等，都属于连续型数据。

3. 按照度量数据所用的测量量表分类

按照度量数据所用的测量量表等级，可以将数据分为定类数据、定序数据、定距数据和定比数据。

定类数据是指用定类量表测量表示的数据，用以表示研究对象所属的类别，没有顺序性、等距性、可加性，不能对数据进行大小比较，更不能对数据进行加减乘除运算。例如，男生用“1”表示，女生用“2”表示，这里的“1”、“2”既无大小之别，也不能参与加减乘除运算。

定序数据是指用定序量表测量表示的数据，表示按大小、轻重、等第等特征依次排列的测量属性，这类数据具有顺序性，但不具备

等距性与可加性，可以进行大小比较，但是不能参与加减乘除运算。例如，学生测验成绩划分为优、良、中、差四个等级，分别用1、2、3、4表示，这些数据具有大小顺序，具有传递性，但不能参与加减乘除运算。

定距数据是指用定距量表测量表示的数据，表示测量遵循统一的单位，相等的点与点之间的距离也是相等的，即测量特征具有顺序性、等距性，可以进行大小比较，也可以参加加减运算，但是不能参与乘除运算(因为定距变量没有绝对零点，例如测验得0分不表示该考生相应的能力也为0)。

定比数据是指用定比量表测量表示的数据，除含有定距数据的特征之外，还有绝对的零点(即表示被测量的属性完全没有)，因此，可以加减乘除四则运算。

一般地，物理测量中所使用的数据大都是定距数据和定比数据，而教育测量中获得的数据以定序数据、定距数据居多。

二、表示测验成绩的数据的基本特点

1. 同一次测验中的分数都是定距数据

教育测验中的分数基本上采用百分制，百分制本质上是一种定序量表，它只是规定了同一类别中数据的大小顺序，并不要求每个排名之间的距离是一样的，如，不要求“20分与10分之间的差别”等于“60分与50分之间的差别”。然而，在教育实践中，人们通常假设同一次测验中“每1分都是等值”的(虽然事实上不等值)，即假设百分制是定距量表，对表示测验成绩的数据进行有关的加减运算。

在教育实践中，为了从测验分数中获取更多的信息，人们还对测验分数进行有关的乘除运算、乘方与开方运算(如计算测验信度、效度等)，实际上把百分制又看成了定比量表，这是违背测量原理的，例如，在一次数学测验中，学生A得80分，学生B得40分，学生A的分数是B的分数的两倍，但是人们不能说学生A的相应数学能力是学生B的2倍，因为，测验的零分并不代表数学能力为零。然而，由于可以解决

部分实际问题，有时候人们有意忽略这种逻辑矛盾。

另外，不同测验的原始分数事实上不等值，如果要把不同测验上的原始分数进行加减运算，理论上应该把这些原始分数转化为标准分数，即转化到同一个单位下的定距量表中，才能进行下一步的处理。但是，在实际操作中，为了快捷地、部分地解决问题，有时候人们也有意忽略这种使用前提，直接进行了不同测验间原始分数的简单相加。

2. 测验分数形式上是客观数据，本质上带有主观特征

教育测验测量的是学生的能力状况，它是一种心理特质，只能采用间接测验的方法通过学生对测验试题的反应情况去推断其心理活动的特点与水平。测验时，所使用的测验卷是由命题者命制，测验效度如何受限于命题者的主观认识与能力，考生答题情况也受限于考生参加考试的身心状态，而表示考生水平的测验分数也受限于评卷教师对评分标准的制定与使用水平。因此，测验分数形式上体现为客观数据，实质上含有很多主观成分，在使用这些数据进行分析推断时，需要慎重考虑，不能武断地给出类似“分数低，则能力就差”的结论。

3. 测验分数刻画的事物特征具有不确定性与模糊性

由于测验分数表示事物特征具有较多主观性，处理测验分数的方式也比较粗糙，因此用测验分数刻画事物特征并不是精确的，它更多带有不确定性与模糊性的特点。这种不确定性与模糊性更基本、更深刻地反映了教育测验的客观现实，因为精确性是相对于某种实际需要而言的，是模糊性被忽略、扬弃大量次要因素后的特例。在处理测验结果、解释测验结果时更需要使用模糊数学的方法进行研究，不能过高苛求测验结果处理的精确性。

第二节　测验分数的组织与表达

无论是学校常规教学中的测验，还是大规模的教育考试，获得

的测验数据都比较多。如果对每一个实测分数作分析处理，不仅需要花费大量的时间与精力，而且效果未必理想。对这些测验数据进行适当组织，可以使得数据中蕴含的规律显现，而且易于理解。

任何数据组织与整理的第一步都是从数据的排序开始的。整理数据时，需要借助一些有效的方法对数据进行分组、排序与对比。

【例3－1】 表3－1是高二(2)班58名学生某次单元测验的成绩，请找出测验成绩的最高分、最低分与众数。测验成绩在分数范围内是否均匀分布？

表3－1　高二(2)班58名学生某次单元测验成绩

76	90	78	77	74	68	60	74	68	94
66	84	81	85	54	91	62	65	75	79
80	93	55	89	76	78	81	80	76	52
75	76	68	80	75	55	87	72	68	57
69	93	57	84	76	69	78	87	37	60
85	76	62	76	48	90	91	73		

显然，直接根据表3－1来回答例3－1的几个简单问题，显得很不容易。但如果掌握了一些快速组织数据的方法，那么解答这几个问题就显得很容易了。

一、茎叶图

茎叶图是一种非常有效的探索数据分布状况的数据分析方法，它的操作非常容易掌握。现在，把表3－1中的数据利用茎叶图表示出来。

如图3－1，竖线左边的数据表示茎，代表每个数据十位上的数字；竖线右边的数据表示叶，代表每个数据个位上的数字。最右边一列的数字表示包含在这一组中的数据个数。

9	0011334	7
8	000114455779	12
7	2344555666666667889	19
6	002256888899	12
5	245577	6
4	8	1
3	7	1

$N = 58$

图 3-1 表 3-1 数据的茎叶图

根据图 3-1，现在可以很容易地回答例 3-1 的问题，最高分是 94 分，最低分是 37 分，众数是 76 分，显然数据的分布并不均匀，绝大多数的数据落在 60～80 之间。另外，人们还可以读出其他的一些有用的信息。

茎叶图的主要优点是保留了数据的所有细节信息，但是，当数据量太大时，这种优点又变成了缺点，比如会导致茎很长或叶子很多。当数据过多时，可以先对数据进行一些折中的处理，比如，如果要画出 0～999 之间的 100 个数据的茎叶图时，可以把每个数据的个位数字截去，茎表示百位上的数字，叶表示十位上的数字，这样得到的修正后的茎叶图保留了数据的主要特征，有效信息也得到较好的保持。

【例 3-2】 表 3-2 是高二(1)班 38 名学生某次单元测验的成绩，请用茎叶图比较同一次单元测验中，该班与例 3-1 中的高二(2)班的成绩情况。

表 3-2 高二(1)班 38 名学生某次单元测验成绩

98	87	82	77	74	90	89	91	95	90
94	77	84	85	85	93	77	84	75	99
91	99	92	80	99	92	79	91	94	75
99	73	85	74	87	93	98	94		

解：图 3-2 是包含了高二(1)、高二(2)两个班级考试成绩的茎叶图，其中茎放在中间，茎左边的叶表示高二(1)班的分数，茎右边的叶表示高二(2)班的分数。

高二(1)班		高二(2)班
9999885444332211100	9	0011334
9775554420	8	000114455779
977755443	7	2344555666666667889
	6	002256888899
	5	245577
	4	8
	3	7

图 3-2 两个班单元测验成绩的茎叶图

从图 3-2 可以很明显地看出，高二(1)班的成绩远远高于高二(2)班的成绩。基本上可以断言，高二(1)班应该属于类似由数学特长生组成的特长班。

需要特别注意的是，使用茎叶图进行两组数据的比较时，需要两组观察数据个数相近或相等。当两组数据数量相差很大时，可能会出现错误的解释。

二、频数分布表

频数即某个数据出现的次数，频数分布即一批数据中各个不同数值所出现的次数的情况。整理一批数据时，除了使用茎叶图外，对数据进行分组归类，考查这批数据在各个等距组内的次数分布情况，并把这种情况用规范的表格表示出来，这就是频数分布表，这种方法在初步理解数据的基本信息时也很有用。表 3-3 是例 3-1、例 3-2 中的两个班某单元测验成绩的频数分布表。

表3-3 两个班某单元测验成绩的频数分布表

分数段	频数	
	高二(1)	高二(2)
[90, 100)	19	7
[80, 90)	10	12
[70, 80)	9	19
[60, 70)	0	12
[50, 60)	0	6
[40, 50)	0	1
[30, 40)	0	1
合计	38	58

与茎叶图相比，频数分布表损失了原始数据的信息，但是呈现数据分布规律时显得很清晰，有时候，这种分析主题更突出。

虽然频数很有用，但如果还想进行更细致的分析，寻找对数据更深层的解释，需要一种简便快捷的数据，例如可以将频数转化为频率(相对百分比)来解决问题。这时，表3-3可以用表3-4的形式呈现。

表3-4 两个班某单元测验成绩的频率分布表

分数段	频数	
	高二(1)	高二(2)
[90, 100)	0.50	0.121
[80, 90)	0.263	0.207
[70, 80)	0.237	0.328
[60, 70)	0	0.207
[50, 60)	0	0.103
[40, 50)	0	0.017
[30, 40)	0	0.017
合计	1.00	1.00

运用表 3-4 来分析两个班的成绩，避免了两个班总人数不同的问题，显得说服力更强。

【例 3-3】 高二(2)班学生 A 考了 80 分，请分析他在班级中所处的位置。

要知道学生 A 的分数是高还是低，需要根据其他学生的分数分布情况来判断。利用茎叶图(图 3-1)可以很清晰地数出学生 A 排在班级的第 17～19 名，处于班级的中上水平。

另外，人们也可以用累计频数(频率)分布表进行分析。如表 3-5，约有 67.2%的分数比学生 A 的分数低，也就是说学生 A 的分数处于 67.2%百分位数点，这种表述比“处于班级的中上水平”更令人信服。

表 3-5 高二(2)班某单元测验成绩的频数分布表

分数段	频数	累计频数	累计频率(%)
[90, 100)	7	58	100
[80, 90)	12	51	87.9
[70, 80)	19	39	67.2
[60, 70)	12	20	34.5
[50, 60)	6	8	13.8
[40, 50)	1	2	3.4
[30, 40)	1	1	1.7

需要注意的是，一个孤立的分数是没有分析意义的，必须把一个百分位数放在相应的组中进行比较才有意义。比如，八年级学生 D 说他在某项数学能力测试中的百分位点是 88，而测验组都是由八年级学生组成，得到这个分数很好，但不出奇；但如果测验组都是由高三学生组成，那么就需要对学生 D 刮目相看了。

三、频数分布直方图

为了更直观、更形象地表达一个频数分布的结构形态及其特

征,人们可以从频数分布表出发,绘制出相应的频数分布直方图。图3-3是根据表3 5绘制的高二(2)班某单元测验成绩频数分布直方图。

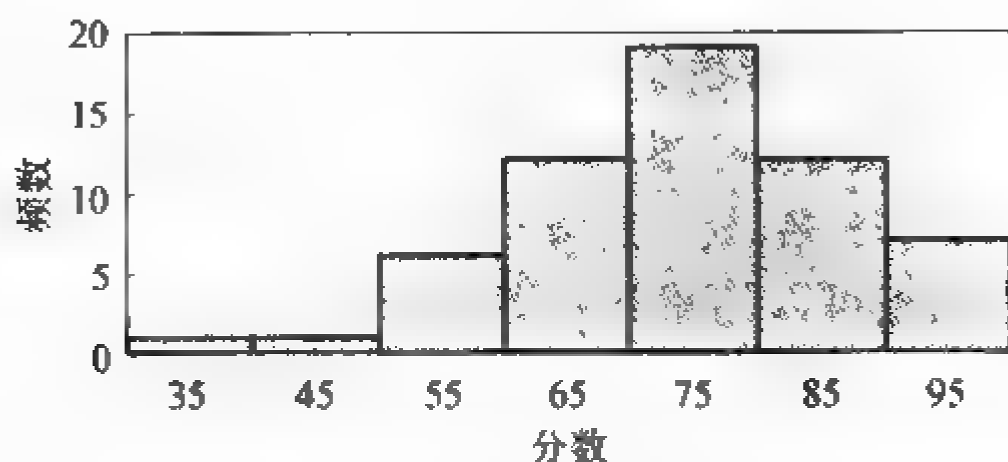


图3-3 高二(2)班某单元测验成绩频数分布直方图

一般地,横轴代表测验分数,其测验分数的下限决定着横轴刻度的起始位置,然后按照适当的比例等间隔地标示出频数分布各组的组中值(图3-3中标出35、45、55、65、75、85、95);纵轴表示频数,按比例等间隔地标出刻度(图3-3中标出0、5、10、15、20即可),其刻度往往从"0"开始。

直方图虽然直观形象,但是也有美中不足,人们不容易准确、快速地了解各组中频数的大小。因此,有人也把各组频数分别标注在各个直方条的顶端,以便阅读。

第三节 测验分数的图表表示

图形和表格是数据整理与分析中不可缺少的部分,它们帮助人们依据研究目的与研究内容组织、总结和解释数据。当使用图表时,数据更容易理解和解释,也更容易让人记住。

一、图形的特点

在第二节中,人们已经接触到一些统计图(如茎叶图、直方图),

好的统计图具有一些共同的特点及其一般形式，这些共同的基本特点能够确保以尽量简单的方式准确地展示数据。

1. 统计图都放置在直角坐标系中(有时候，直角坐标系没有直接显示出来)。

2. 横轴(x 轴)用来表示自变量，纵轴(y 轴)用来表示因变量。分析时，用自变量来预测、描述或解释因变量的变化。

3. 横轴(x 轴)与纵轴(y 轴)都有明确的标识。所有的统计图都必须包括测量的尺度、变量的名称以及图示说明。

4. 横轴(x 轴)与纵轴(y 轴)的单位长度比按照黄金分割作画。即图形的宽度约为高度的 1.7 倍，这样的图形最美观。

5. 纵轴(y 轴)标示连续数据，没有充分的理由，不能断开纵轴(y 轴)。

6. 统计图中只给出和数据有关的信息。一般情况下，不需要在图中的点、线上标注具体数值。因为图形的目的是为了让读者有一个快速直观的印象，而不是数据的具体内容。如果需要给出具体的细节信息，那么使用表格。

二、常用统计图

1. 条形图

条形图是用宽度相同的长条表示各个统计对象之间的数量关系，它在考试数据分析中经常用到。它具有以下三个独特的特征。

(1) 横轴表示的是定类数据或定序数据。

当数据代表的是不同类别或者不同等级的情况时，就使用条形图。

(2) 长条与长条之间是间断的。

由于横轴表示的是定类数据或定序数据，因此一个长条代表一个类别，表明这些类别是离散的。

(3) 纵轴可以表示频数、百分数，或其他描述性统计量。

【例 3-4】 表 3-6 是一道“四选一”型选择题的各选择支答题

情况统计，请用条形图表示该题各个选项情况。

表3-6 一道选择题答题情况统计

选A率	选B率	选C率	选D率	未选率
12.79%	19.26%	61.18%	5.51%	1.24%

解：如图3-4，其中横轴表示选择题选择支的分类，未选率用NO表示；纵轴表示的是选择各个选项的百分数。

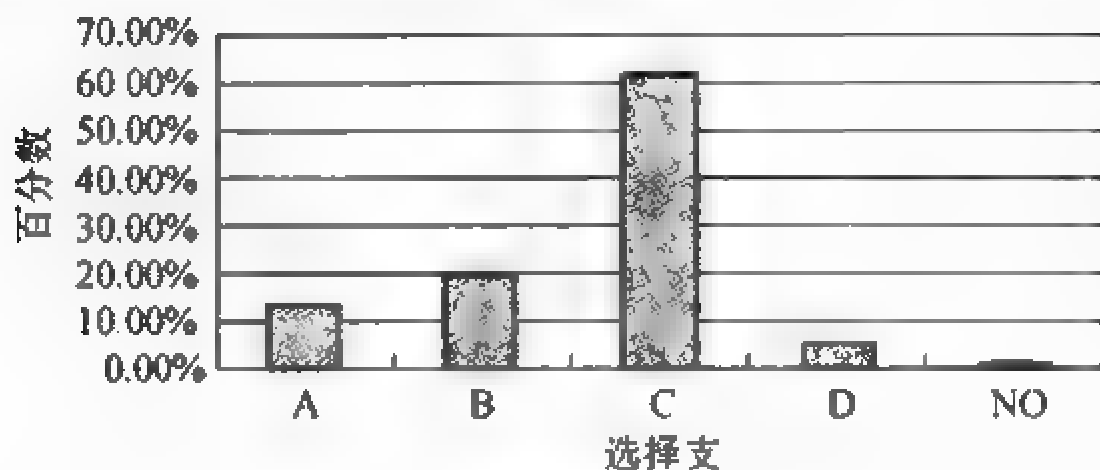


图3-4 一道选择题答题情况条形图

2. 直方图

直方图与条形图的最主要的区别有两点：

(1) 横轴表示的数据类型不同。

直方图的横轴表示的数据属于定距数据或定比数据，数据代表的分数呈现出由左向右、从低值到高值的连续变化。

(2) 长条呈现形式有别。

直方图也用长条表示数据，每个长条代表一个类区间，类区间就是有确定上限与下限的数值取值范围。由于直方图的横轴代表的是连续变量，因此，长条与长条间没有缝隙。

除了图3-3所示的频数分布直方图外，常见的还有频率分布直方图，即从各个小组数据在样本容量中所占比例大小的角度，来表示数据的分布规律。图3-5是根据表3-4中的数据绘制出的高二(2)班某单元测验成绩频率分布直方图。

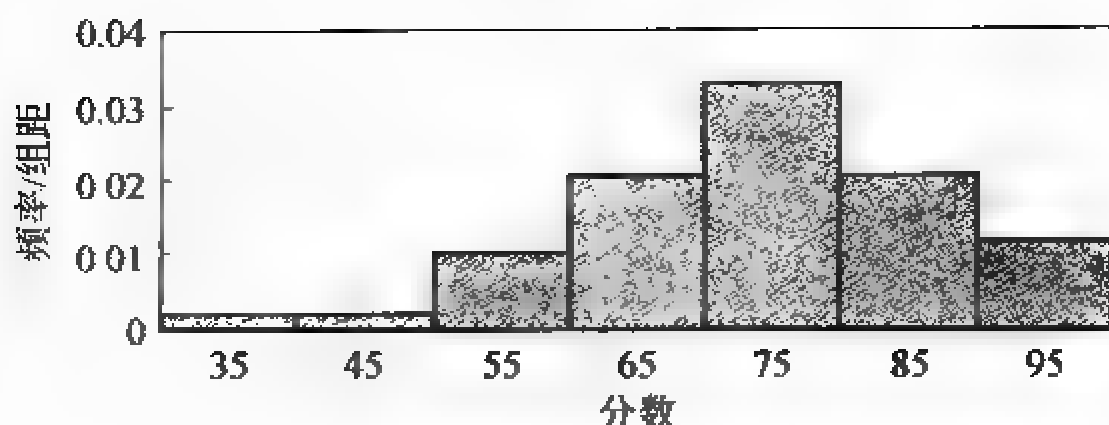


图 3－5 高二(2)班某单元测验成绩频率分布直方图

图 3－5 中，测验分数共分为 7 组，组距是 10(分)，横轴表示测验分数，纵轴表示频率/组距。由于每个长条(小长方形)的面积＝组距×频率/组距＝频率，即频率分布直方图是以面积的形式反映数据落在各个小组的频率的大小，所以各个长条面积总和等于 1。

需要注意的是，同样一组数据，如果组距不同，横轴、纵轴的单位不同，得到的直方图形状也会不同。不同的形状给人以不同的印象，这种印象有时会影响人们对总体的判断。

3. 茎叶图

上节已经详细介绍了茎叶图的使用与它的种种优点，它既是用于了解数据的工具，也是展示数据的有效方法。茎叶图除了前面介绍的呈现形式外，还可以用类似直方图的形式展示数据，如图 3－6 所示。

				9		
				8		
				8		
				8		
				7		
				6		
				6		
			9	6	9	
			9	6	7	
			8	6	7	
			8	6	5	
			8	6	5	
			8	5	4	4
			6	5	4	3
		7	5	5	1	3
		5	2	4	1	1
		5	2	4	0	1
		4	1	3	0	0
7	8	2	0	2	0	0
3	4	5	6	7	8	9

图 3－6 图 3－1 中数据对应的茎叶图

图 3－6 中，数字替代了直方图中的长条，既保留了直方图的特征，又保留了每个学生的原始成绩信息，因此，任何人都很容易看到自己相对于其他学生的成绩情况。

在学校中，一个班级的人数在 40～60 人之间，制作茎叶图很方便，

教师用茎叶图向学生公布成绩，既可以保护学生的分数隐私，又可以让学生一看就清楚知道自己在全班中所处的位置。因此，建议教师多使用茎叶图。

4. 散点图

散点图是用平面直角坐标系上点的散布图形来表示两种变量之间的相关性与联系模式。散点图适合于描述二元变量的观测数据，在探索变量之间的变化规律方面有独特的作用。

图 3-7 与图 3-8 分别表示的是初三(3)班某次数学测验的测验总分与主观题(解答题)、客观题(选择题和填空题)得分情况的关系。由图 3-7 可知，主观题得分高则测验总分高，二者间具有明显的线性关系；图 3-8 则表明，总体上而言，客观题得分高则测验总分高，但是仍有不少学生客观题得分高，但测验总分并不高。出现这

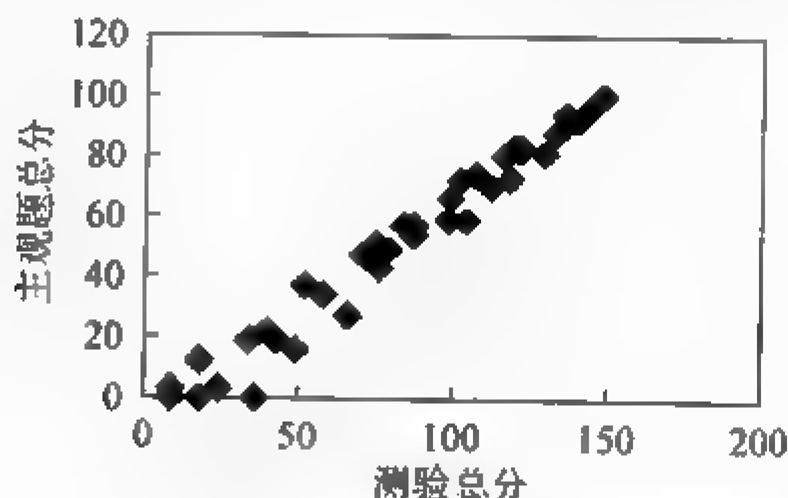


图 3-7　初三(3)班某测验总分与主观题得分散点图

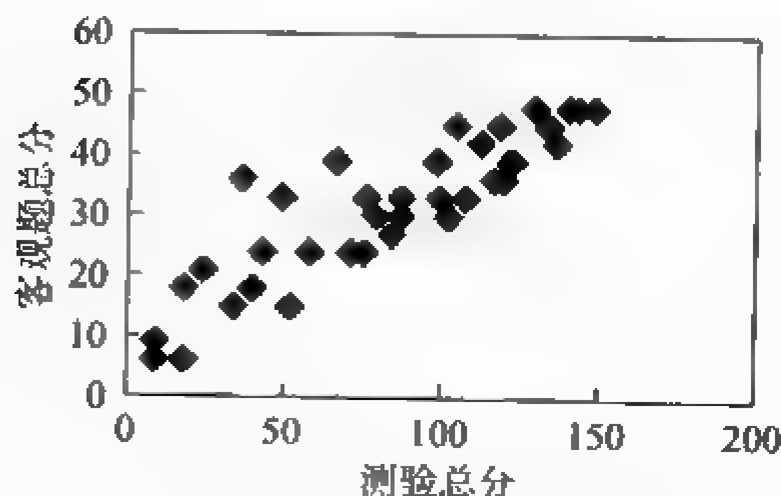


图 3-8　初三(3)班某测验总分与客观题得分散点图

种情况，需要结合考生的具体答卷情况，再进行深入分析。

5. 折线图

折线图是以起伏的折线来表示数据发展变化与演变趋势的统计图，适用于描述某种事物在时间序列上的变化趋势，也适用于比较不同样本或总体在同一个研究对象上的变化特征及其相互联系，因此是常用且有效的测验数据展示方法。

图3-9非常清晰地展示出五名三年级学生在上学期学习中7次常规单元测验成绩之间的差异，显然，在五名学生中，学生E的成绩比较稳定，而且分数位居高位；学生C的成绩相对较弱，波动幅度也较大；而成绩波动最大的是学生B，需要针对性地了解该生的学习习惯与学习特长。

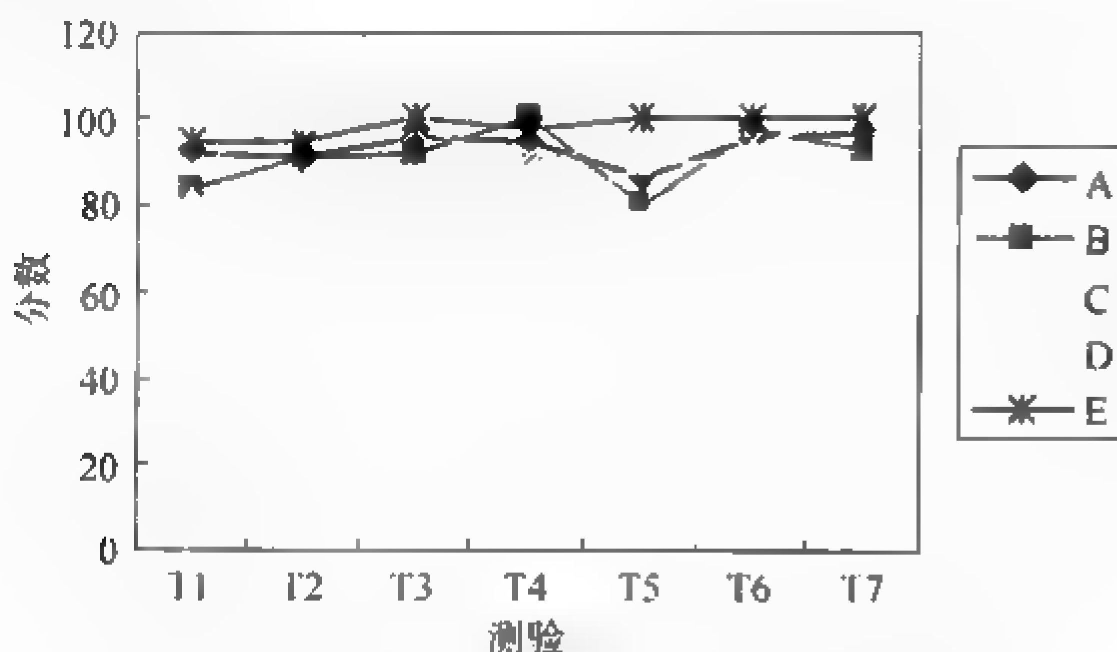


图3-9 五名三年级学生上学期7次测验比较折线图

折线图提供的这些信息对教师了解学生的学习变化情况非常有用，能够指导教师更好地落实因材施教，分类指导。

三、统计表制作的规则

尽管统计图是组织和表示数据的一种非常有用的工具，它能够迅速地反映出数据的分布与特征，但是很多情况下，还需要展示具体的数据。制作一个有用且容易阅读的表格与作图一样重要。与制作好的统计图一样，制作好的、易于读懂的统计表也有一些基本

的规则。艾伦伯格(1977)确定了制作一个好表格的六个基本要求。

1. 应该按照测量精度报道数据

如果报道的数字有小数部分，一般情况下保留两位小数，也可以根据测量精度提高或降低保留位数的个数。例如，计算的实际结果是215.324689，一般表示为215.32。

2. 一般应表示出行和列的均值、和值

均值、和值等概括性统计量可以帮助读者掌握数据的一般趋势以及不同数据组之间的差异。

3. 最重要的数据按列表示

一般认为，在垂直方向上浏览数据比水平方向上浏览数据更容易，为了更好地表达与交流数据，制作表格时，把最重要的数据按列排列。

4. 一般应将数据按照从小到大或从大到小的顺序排列

对数据进行排序后，读者更容易发现数据中的极值，掌握数据的分布范围。

5. 行间距、列间距适当偏小

行间距与列间距只要能够保证数值清晰即可。另外，数值间的距离应该比较接近，这样有利于对相邻行和相邻列之间的数据进行比较。

6. 只有在必须呈现具体数值的情况下才使用表格

由于数据表占用的空间比较大，而且数据比较单调、枯燥，因此，只有必须要用表格或没有其他替代方式时，才采用数据表[①]。

另外，表格中的线条不宜过多。顶线、底线、隔开列名称与数字的横线、隔开行名称与数字的纵线是四种基本线条，其余线条应尽量减少，表的左右两侧不要用纵线封闭。

---

① 王星译，Richard P. Runyon 等，行为统计学基础(第9版)，中国人民大学出版社，2007，134－135页。

四、常用统计表

根据统计表结构的简繁程度，可以将常用统计表分成简单表、分组表、交互表三种。

1. 简单表

即只列出研究对象的名称、地点、统计指标名称的统计表，如表3-7所示。

表3-7 初三年级各班学生人数

班级	1班	2班	3班	4班	5班	合计
人数	45	43	46	42	47	223

2. 分组表

分组表即只按照一个标志分组统计数据，例如第二节中介绍的表3-3、3-4、3-5都是只按照测验成绩这个标志将成绩分成若干个组，对数据进行统计。

3. 交互表

按照两个或两个以上标志分组统计数据的表格称为交互表。表3-8是按照班级、性别、学期成绩等级等三个标志进行分组的，这种统计表也称为三项表。

表3-8 某年级学期成绩等级评定表

班级	优		良		中		差		合计
	男	女	男	女	男	女	男	女	
1班	7	6	10	12	3	4	1	1	45
2班	6	7	11	11	4	3	1	1	43
3班	8	7	10	11	4	3	2	1	46
合计	21	20	31	34	11	10	4	3	134

第四节　测验分数的概括

在分析测验成绩时，用茎叶图、频数分布表、频数分布直方图等初步整理数据的方法，对数据进行列表、图示，可以对数据的分布特征有一定直观而形象的概要了解。但是，如果要对这些数据所蕴含的规律性做更进一步的推论和更好的了解，还需要计算出一些有代表性的数据，对变量所蕴含的规律性做出更简洁的数量化描述，对其频数分布特征做出定量刻画。

一、集中量数

描述数据集中趋势的统计量称为集中量数。不同的集中量数都是描述统计量，都是用来总结、描述数据的集中情况或频繁出现情况。常用的集中量数有三个：平均数、中位数和众数。

1．平均数

在处理测验卷的数据时，人们经常遇到的平均数是算术平均数，它等于得分总和除以得分个数，用公式表示如下：

$$\bar{x}=\frac{x_1+x_2+\cdots+x_n}{n}=\frac{1}{n}\sum_{i=1}^{n}x_i。\qquad(3.1)$$

其中，$\bar{x}$表示算术平均数，x_i 表示每个得分，n 表示数值个数。

平均数是最常用的一个统计量，它具有以下几个重要性质。

(1) 平均数偏差之和等于零。

用公式可以表示为

$$\sum_{i=1}^{n}(x_i-\bar{x})=0。\qquad(3.2)$$

也就是说，平均数在它两边所有数值距离（偏差）达到均衡。在许多方面，平均数和跷跷板的平衡点很相似。

(2) 平均数对极端值很敏感。

一组数据中，如果某个数据特别小(或特别大)，那么这组数据的平均数就会向这个特别的数据靠近，以求得均衡。因此，平均数受组中数据极端取值影响很大。

(3) 平均数偏差的平方和最小。

一组数据中，平均数偏差的平方和比其他任何点的偏差平方和都小，这个性质在统计学中经常遇到，也称为“最小二乘”，即平均数表示自身和一组数据中数值有最小的平均偏差。

很多时候，需要将若干组数据加以合并，并计算合并后的平均数，这种情况下，就需要计算加权平均数，其中各组数据的个数作为计算的权重。

【例 3－5】 初三年级某次数学竞赛各班成绩如表 3－9 所示，求该年级数学竞赛平均分。

表 3－9 初三年级某次数学竞赛各班平均成绩表

班级	1 班	2 班	3 班	4 班	5 班	合计
人数	45	43	46	42	47	223
平均分	83	78	80	86	79	?

解： $\bar{x} = \sum(f_i \cdot \bar{x}_i)$

$$= \frac{45}{223} \times 83 + \frac{43}{223} \times 78 + \frac{46}{223} \times 80 + \frac{42}{223} \times 86 + \frac{47}{223} \times 79$$

$$\approx 81.14。$$

即初三年级数学竞赛平均分为 81.14。

平均数具有反应灵敏、确定严密、简明易解、计算简便而且能够做进一步代数运算等优点，因此使用非常广泛。但是，当一组数据中存在较多的极值，或数据分布明显偏态时，使用平均数容易产生误导，应引起注意。

2. 中位数

中位数就是将一组数据分为两半的数值，有一半的数值大于中

位数，有一半的数值小于中位数，即中位数是50%的百分位点。

把一组数据按照由小到大、或由大到小的顺序排列，设这组数据的个数是 n，当 n 是奇数时，中位数的位置就是 $\frac{n+1}{2}$，即第 $\frac{n+1}{2}$ 个数是这组数据的中位数；当 n 是偶数时，中位数的位置就是 $\frac{n}{2}$ 和 $\frac{n}{2}+1$，即第 $\frac{n}{2}$ 和 $\frac{n}{2}+1$ 个数是这组数据的中位数。

注意，当 n 是偶数时，中位数有两个。另外，也要注意区分中位数的位置与中位数，两者不同，不要混淆。

由于中位数主要由数据排序后的位置决定，它不受极端值的影响，因此，如果知道数据分布明显是偏态时，考虑优先使用中位数。但中位数对存在极端高或极端低的数据不敏感，使用时也需要注意。

3. 众数

众数就是一组数据中出现频率最高的数值，它最容易得到，因为它是观察得出的，而不是计算得到的。作为一个描述统计量，众数并不常用，因为它不够精确，信息量有限，而且容易产生误导。当数据的度量是采用定序量表、定距量表、定比量表时，一般都采用平均数或中位数；当数据的度量是采用定类量表时，可以使用众数。

二、差异量数

频数分布中数据间彼此差异的程度称作数据的离中趋势，离中趋势反映了频数分布的离散程度。描述离中趋势的统计量称为差异量数。最常见的差异量数是：极差、最大值与最小值、方差和标准差。

1. 极差

一组测验分数的极差就是这组分数中最大值与最小值的差。例如，一组分数中，最高分是98分，最低分是35分，则极差就是63分。

极差的计算非常简单，它刻画了数据的波动范围，很有意义。如果极差比较大，说明考生的考分差异较大；如果极差比较小，则说明考生的考分比较集中；若再结合测验平均分进行对比，就能很容易地了解全体考生相应知识技能掌握水平的高低。但是，由于极差只考虑了最高分与最低分，如果数据中存在一个异常值，那么极差反映的散布程度可能就非常大，而去除这个异常值后，可能数据分布非常集中，这样就容易产生误导。

2. 最大值与最小值

虽然极差可以描述一组测验分数的散布程度，但是它并没有具体刻画这组测验分数的最大值与最小值。而在教育测验中，特别需要关注测验分数的极端值，因此需要知道测验分数的最值。

在分析最大值与最小值时，特别需要关注最值附近的异常点。如果极差不大，最值附近数据较多，则表明数据分布比较正常。如果极差很大，而最值附近的数据不多，那么需要考虑最值附近的这些数据产生的原因。

3. 方差与标准差

在学校教育测验中，人们往往把考生成绩看成一个样本，因此，在分析分数分布的离散程度时，研究的都是样本方差与样本标准差。

样本方差的计算公式为

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2。\tag{3.3}$$

样本标准差的计算公式为

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}。\tag{3.4}$$

方差与标准差是表示一组数据离散程度的最好指标，其值越大，说明数据分布的离散程度越大；其值越小，说明数据比较集中，离散程度越小。当然，如果一组数据中存在极端值的话，那么方差

与标准差对数据离散程度的刻画可能出现偏差，这点需要注意。

方差往往与平均数结合起来一起使用，用于描述一组数据的全貌。

三、分数分布的形状

描述一组测验分数分布的整体形状时，经常使用的描述统计量是偏度与峰度。

1. 偏度

当一组测验分数中的一端存在极端值时，称这组测验分数的分布为有偏分布，那些离平均数和其他数很远的极端值也称为异常点。如果异常点大于平均数，则称分布为正偏；如果异常点小于平均数，则称分布为负偏。

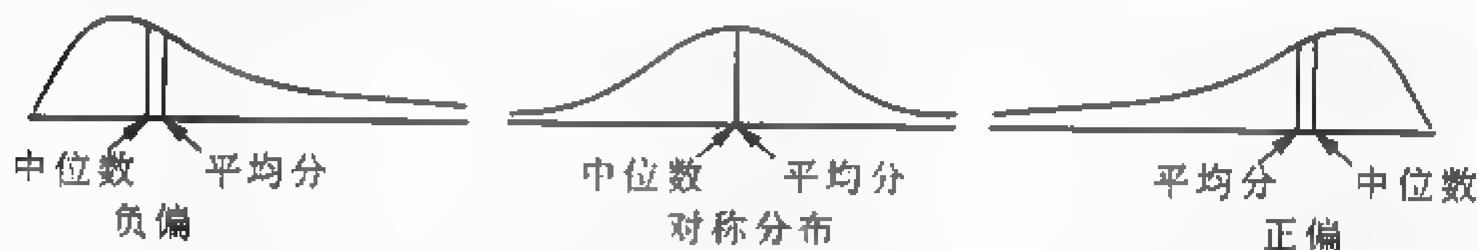


图 3-10 各种分布中中位数、平均分与偏度的关系①

由图 3-10 可知，当平均数大于中位数时，分布正偏；当平均数小于中位数时，分布负偏；当平均数等于中位数时，分布是对称的。因此，当分布有偏时，最好不要用平均数表示数据的集中程度，而用中位数比较适宜。

除了用上述方法来判断分布的偏度外，在统计学上偏度的理论计算公式如下：

$$s^3 = \frac{\frac{1}{n}\sum(x_i - \bar{x})^3}{\left[\frac{1}{n}\sum(x_i - \bar{x})^2\right]^{1.5}}。 \qquad (3.5)$$

---

① 雷新勇，考试数据的统计分析和解释，上海：华东师范大学出版社，2007，44—52 页。

其中，s^3 表示分布的偏度，x_i 是分布中的数据，$\bar{x}$为平均数。

由于公式 3.5 计算使用比较繁琐，因此，在实际应用时，人们常用皮尔逊偏度系数来估计分布偏度，近似计算公式如下：

$$\hat{s}^3 = \frac{3(\bar{x} - M_e)}{s}。 \tag{3.6}$$

其中，$\hat{s}^3$ 表示偏度估计值，即皮尔逊偏度系数；$\bar{x}$为平均数，M_e 是中位数，s 是标准差。

当$\hat{s}^3$ 为正时，分布为正偏；当$\hat{s}^3$ 等于零时，分布为对称的；当$\hat{s}^3$ 为负时，分布为负偏。在实际问题中，当$\hat{s}^3$ 值在±0.50 之间时，就可以把分布看成是对称的。

2. 峰度

峰度是描述数据分布的另一个常见的统计量，它表明数据是否集中在均值附近，或是否有许多极端值且分布在较大的范围。

峰度的理论计算公式如下：

$$s^4 = \frac{\frac{1}{n}\sum(x_i - \bar{x})^4}{\left[\frac{1}{n}\sum(x_i - \bar{x})^2\right]^2}。 \tag{3.7}$$

其中，s^4 表示分布的峰度，x_i 是分布中的数据，$\bar{x}$为平均数。

公式 3.5 与公式 3.7 形式很接近，事实上，偏度又称为平均数的三阶矩，峰度又称为平均数的四阶矩。

同样地，峰度对极端值也很敏感，与偏度类似，人们也有计算峰度的近似计算公式：

$$\hat{s}^4 = 3 + \frac{Q_3 - Q_1}{2(P_{90} - P_{10})}。 \tag{3.8}$$

其中，$\hat{s}^4$ 表示峰度估计值，Q_3、Q_1 分别表示第三和第一四分位数，P_{90}、P_{10}分别表示 90%和 10%分位数。

峰度与偏度的计算都可以用计算机来完成，教师在进行数据分析时不需要背这些公式。

第五节　EXCEL 与 SPSS 软件应用实例

从教育测验中获得的大量测验分数，经过分组、编表、作图等统计方法归纳、整理后，以直观形象的方式体现出分布特征；然后用集中量数（平均数、中位数、众数等）表示测验分数的集中程度，用差异量数（极差、方差与标准差等）刻画测验分数的离散程度；再用偏度值和峰度值反映分数分布的形态，这些描述性统计量的计算、图表的绘画都可以用 EXCEL 和 SPSS 软件来完成。

一、对测验分数进行初步整理

【例 3－6】　以表 3－1 与表 3－2 中高二(2)班与高二(1)班某次数学单元测验成绩为例，对两个班的测验分数进行分组，并计算相应的频数。

解法 1　用 EXCEL 软件，共分四步完成。

第一步，将两个班数学单元测验成绩输入 EXCEL 工作簿的工作表区内。

第二步，求出两个班测验分数的最大值与最小值。

最大值、最小值的计算采用函数“MAX”与“MIN”，在单元格 D3 中键入“＝MAX(C3:C98)”，按 Enter 键，返回值 99 就显示在单元格 D3 中。D3 表示的是两个班 96 名学生数学单元测验成绩的最大值，如图 3－11。在单元格 D4 中键入“＝MIN(C3:C98)”，按 Enter 键，返回值 37 就显示在单元格 D4 中。D4 表示的是两个班 96 名学生数学单元测验成绩的最小值。

第三步，将数据分组，确定各组的分点。

根据测验成绩的最大值与最小值，先将测验分数确定分成 7 个组，即把 37 至 99 分划分为 7 组，组与组之间的分点分别是 39、49、

D3 =MAX(C3:C98)

	A	B	C	D
1	高二年级1、2班			
2	班级	学号	成绩	最值
3	1	1	98	99
4	1	2	94	37
5	1	3	91	
6	1	4	99	
7	1	5	87	
8	1	6	77	

图 3－11

59、69、79、89，并约定若分数 $x \leqslant 39$，则 x 落在[30，39]内；若 $x \leqslant 49$，则 x 落在[40，49]内；……；若 $x > 89$，则 x 落在[90，99]内。

将分组与分点按列输入到工作表区内，如图 3－12。

I16

	A	B	C	D	E	F
1	高二年级1、2班					
2	班级	学号	成绩	最值	分组	分点
3	1	1	98	99	[30, 39]	39
4	1	2	94	37	[40, 49]	49
5	1	3	91		[50, 59]	59
6	1	4	99		[60, 69]	69
7	1	5	87		[70, 79]	79
8	1	6	77		[80, 89]	89
9	1	7	99		[90, 99]	
10	1	8	73			

图 3－12

第四步，统计两个班各分数段的频数。

频数的计算采用函数"FREQUENCY"。选定"1 班频数"单元格下的 7 个单元格 G3 至 G9，单击"常用"工具栏中的按钮[fx]，出现"插入函数"对话框。如图 3－13，在"选择类别"后的空格中选中"统计"类，在"选择函数"栏目下选中"FREQUENCY"，然后单击"确定"按钮，出现公式选项板"函数参数"对话框，如图 3－14。在"Data_

array"后的空格中输入"C3:C40"(即1班38名学生的成绩),在"Bins_array"后的空格中输入分点"F3:F8",按Ctrl+Shift键的同时,按"确定"按钮,返回值0、0、0、0、9、10、19就显示在单元格G3至G9中。G3至G9表示的是1班38名学生数学单元测验成绩各分数段的人数。

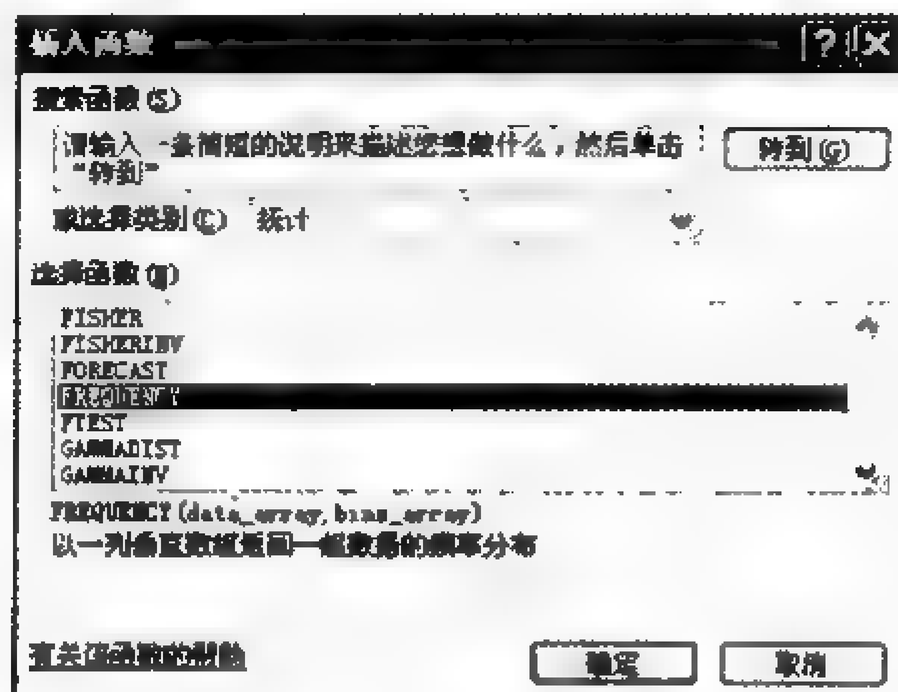


图 3-13

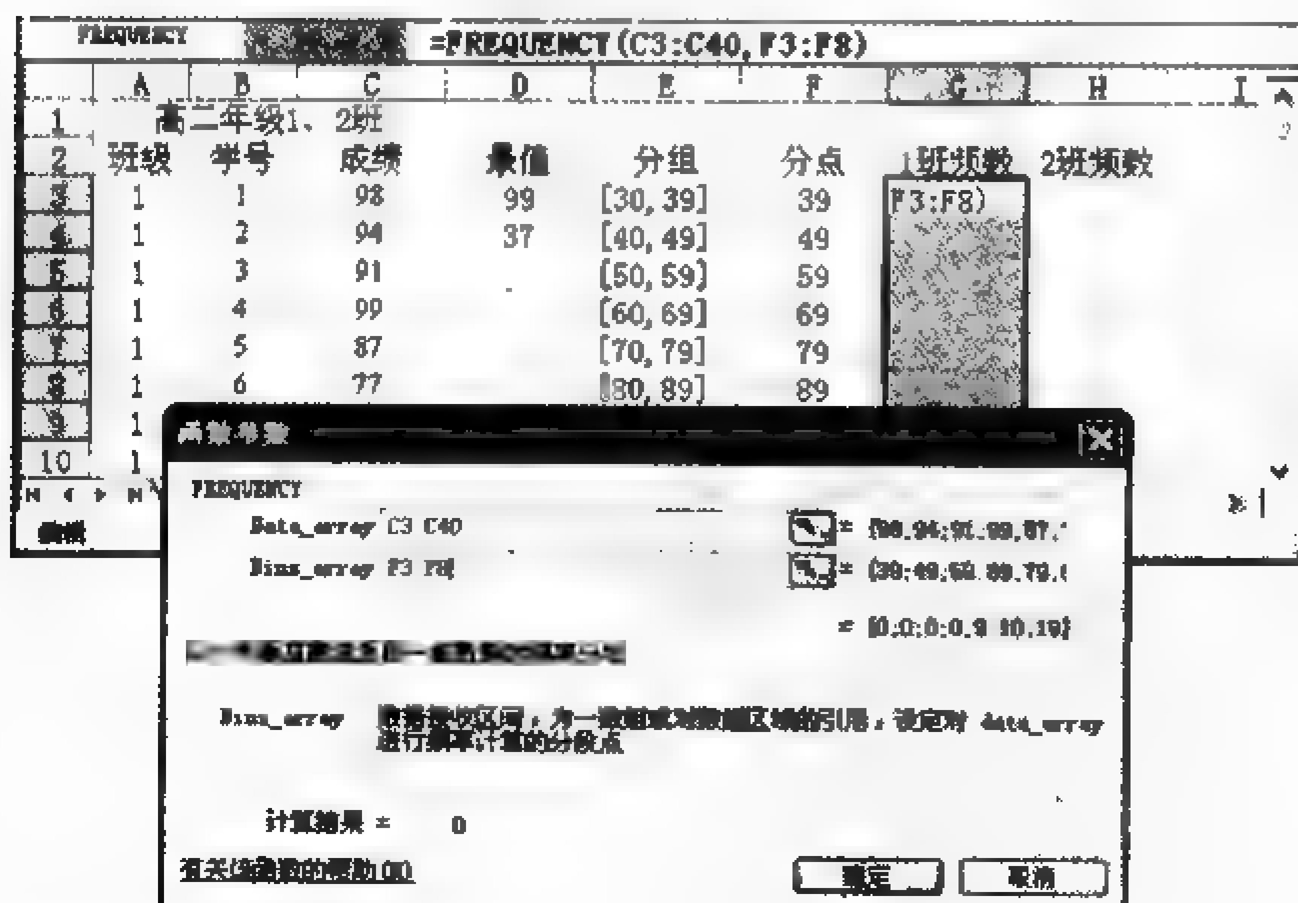


图 3-14

用类似的方法，可以统计出2班58名学生数学单元测验成绩各分数段的人数，如图3－15所示。

H3 {=FREQUENCY(C41:C98,F3:F8)}

	A	B	C	D	E	F	G	H
1	高二年级1、2班							
2	班级	学号	成绩	最值	分组	分点	1班频数	2班频数
3	1	1	98	99	[30,39]	39	0	1
4	1	2	94	37	[40,49]	49	0	1
5	1	3	91		[50,59]	59	0	6
6	1	4	99		[60,69]	69	0	12
7	1	5	87		[70,79]	79	9	19
8	1	6	77		[80,89]	89	10	12
9	1	7	99		[90,99]		19	7
10	1	8	73					

图3－15

解法2 用SPSS软件，共分三步完成。

第一步，将两个班数学单元测验成绩导入SPSS数据编辑器的工作表区内，如图3－16。

	班级	学号	成绩
1	1	1	98
2	1	2	94
3	1	3	91
4	1	4	99
5	1	5	87
6	1	6	77
7	1	7	99
8	1	8	73

图3－16

第二步，把两个班测验分数进行分组。

如图3－17，首先执行【转换】/【可视化分段】程序，出现“可视化分段”对话框。在图3 18中，将左边“变量”下方框中的“成绩”导入到右边“要进行分段的变量”下的方框中。点击“继续”按钮，出现“可视化分段”主对话框，如图3－19。

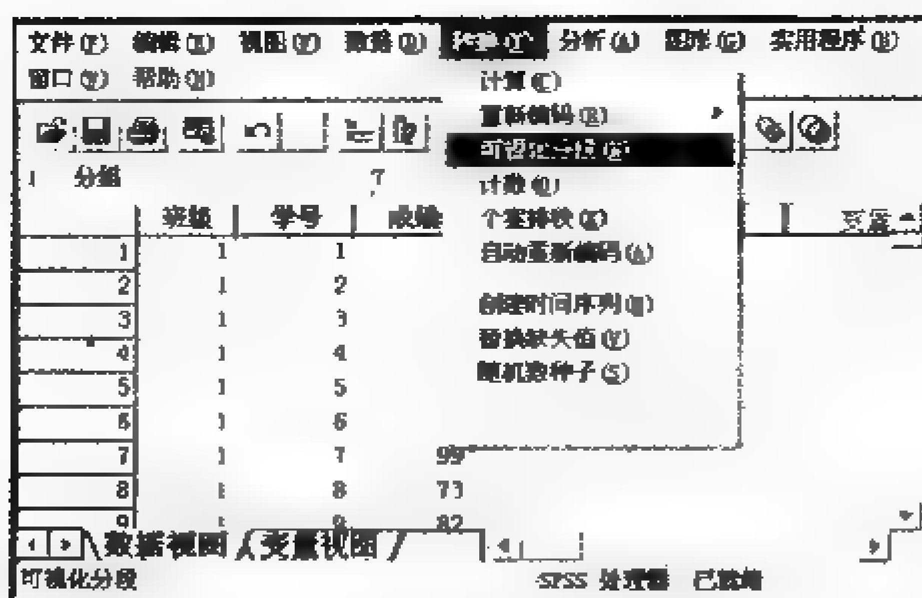


图 3－17

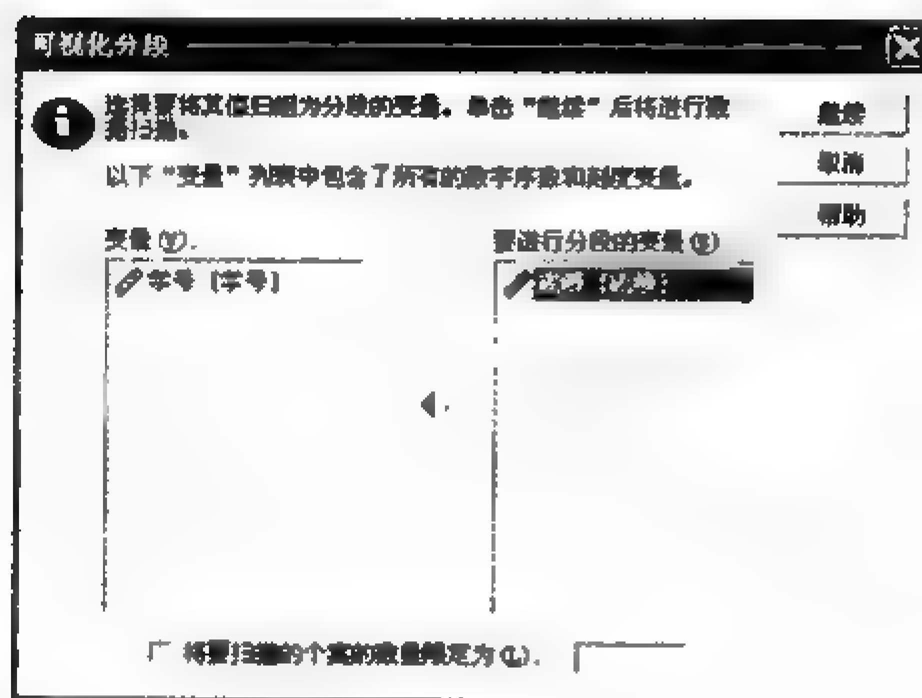


图 3－18

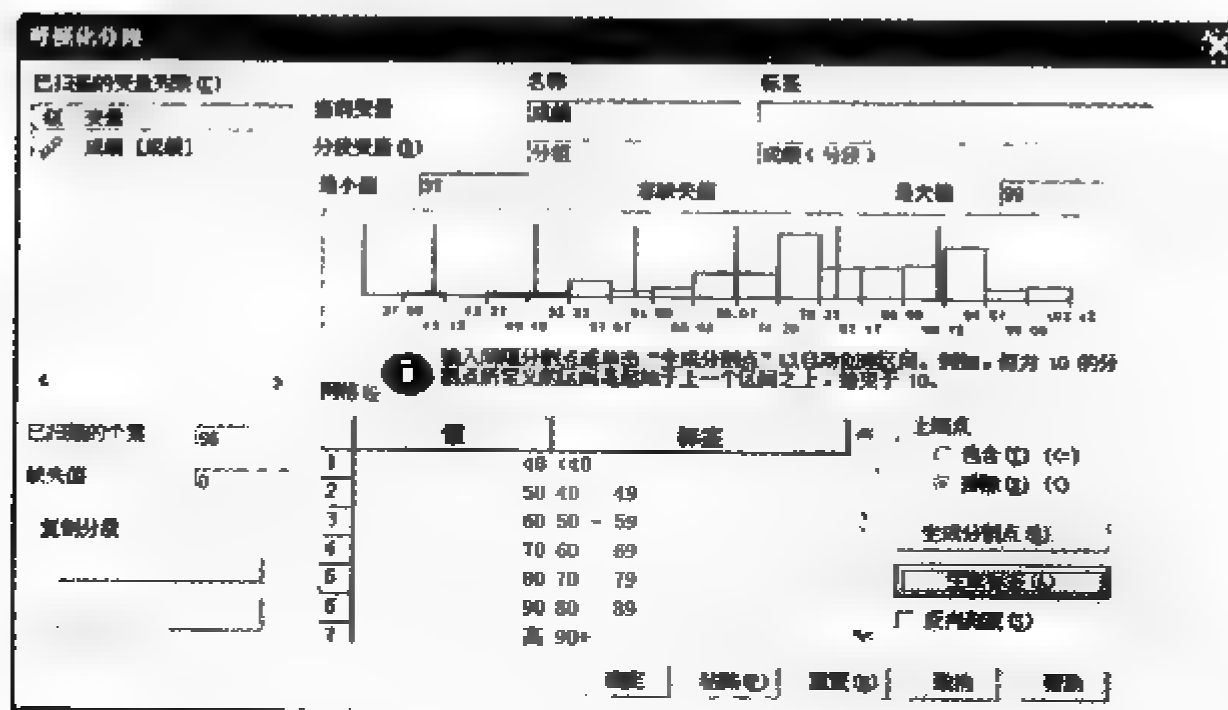


图 3 19

在“可视化分段”主对话框右边“已扫描的变量列表”栏目下选中变量“成绩”，则右边当前变量的名称显示为“成绩”。

在“分段变量”后名称空格中输入“分组”，作为分组后变量的名称。

这时，主对话框的左边显示“已扫描个案”为 96，“缺失值”为 0，表示两个班共有 96 名学生，全部成绩均有效；右边中上部“最小值”显示为 37，“最大值”显示为 99，表示两个班测验成绩的最低分为 37 分，最高分为 99 分。

在“网格”栏目下方有两栏，左边为（分组上限）“值”区，右边为“标签”（即分组的区间表示）。本例中，最低分为 37 分，最高分为 99 分，故将 30～40 分定为第一组，依此类推，每组组宽为 10 分。因此，在“值”区下方第一行输入 40，第二行输入 50，……。然后单击“生成标签”按钮，软件自动在“标签”栏目下生成每一组的标识。

在“上端点”栏目下选中“排出”项，然后单击“确定”按钮。软件自动将数据分组，并将分组结果保存在变量“分组”中，如图 3－20。

文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 分析(A) 图形(G) 实用程序(U) 窗口(W) 帮助(H)

1 : 分组 7

	班级	学号	成绩	分组
1	1	1	98	7
2	1	2	94	7
3	1	3	91	7
4	1	4	99	7
5	1	5	87	6
6	1	6	77	5
7	1	7	99	7
8	1	8	73	5
9	1	9	82	6
10	1	10	84	6
11	1	11	92	7
12	1	12	85	6

数据视图 / 变量视图

SPSS 处理器 已就绪

图 3－20

第三步，统计两个班测验成绩各组的频数。

如图 3－21，首先执行【分析】/【表】/【频率表】程序，出现“频率

表格”对话框，如图3－22。将分组变量“分组”导入右边“频率”栏目下的方框中，将“班级”导入右边“子组——在每个表格中”栏目下的方框中。

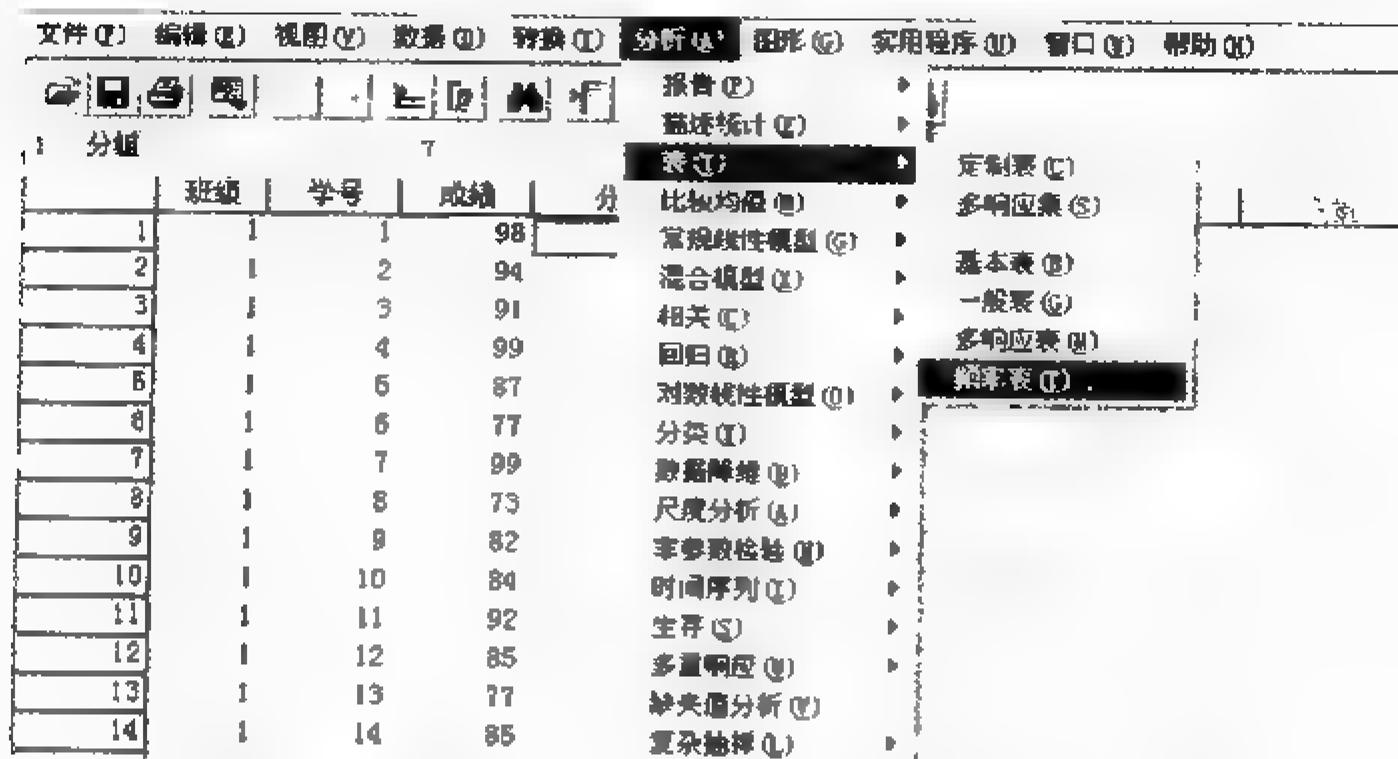


图3－21

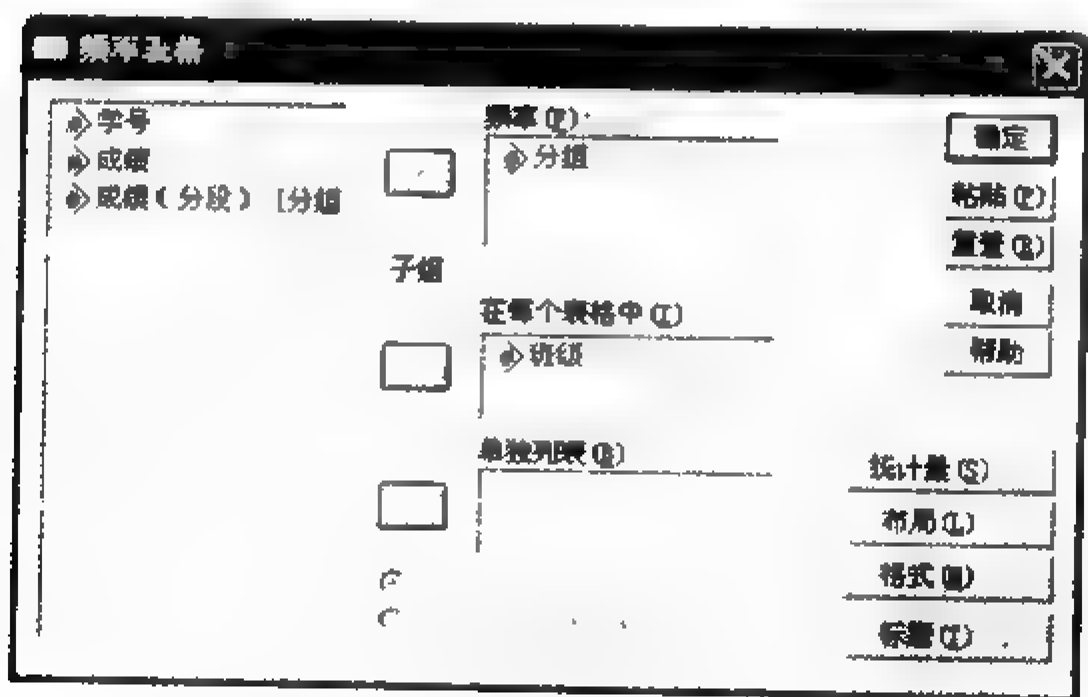


图3－22

单击“统计量”按钮，出现“频率表：统计量”子对话框，如图3－23。选中“计数”栏目下“显示”前的方框(表示显示频数)；选中“百分比”栏目下“显示”前的方框(表示显示频率)。单击“继续”按钮，返回到“频

率表格”。软件自动制作频数分布表，结果如表3－10。

图3－23

单击图3－22中的“确定”按钮，软件自动制作频数分布表，结果整理如表3－10。

表3－10　高二年级1、2班数学成绩频数分布表

	高二(1)		高二(2)	
	频数	百分比(%)	频数	百分比(%)
＜40	0	0	1	1.7
40～49	0	0	1	1.7
50～59	0	0	6	10.3
60～69	0	0	12	20.7
70～79	9	23.7	19	32.8
80～89	10	26.3	12	20.7
90＋	19	50.0	7	12.1

二、计算描述性统计量

【例3－7】 以表3－1与表3－2中高二(2)班与高二(1)班某次数学单元测验成绩为例，用描述性统计量(如平均分、中位数、众数、极差、方差与标准差、偏度与峰度等)来反映两个班的测验分数。

解法1 用EXCEL软件，分两步完成。

第一步，将两个班数学单元测验成绩输入EXCEL工作簿的工作表区内。

第二步，运用统计函数依次求出每个统计量。

1. 计算平均数

平均数的计算采用函数"AVERAGE"，如图3－24，在单元格E2中键入"＝AVERAGE (C2:C39)"，按Enter键，返回值87.39474就显示在单元格E2中。E2表示的是高二(1)班38名学生数学单元测验成绩的平均分。在单元格F2中键入"＝AVERAGE (C40:C97)"，按Enter键，返回值73.87931就显示在单元格F2中。F2表示的是高二(2)班58名学生数学单元测验成绩的平均分。

2. 计算中位数

中位数的计算采用函数"MEDIAN"，其计算步骤与计算平均数类似。如图3－24，在单元格E3中键入"＝MEDIAN(C2:C39)"，按Enter键，返回值89.5就显示在单元格E3中。E3表示的是高二(1)班38名学生数学单元测验成绩的中位数。在单元格F3中键入"＝MEDIAN(C40:C97)"，按Enter键，返回值76就显示在单元格F3中。F3表示的是高二(2)班58名学生数学单元测验成绩的中位数。

3. 计算众数

众数的计算采用函数"MODE"，其计算步骤与计算平均数、中位数类似，计算过程略。

4. 计算极差

极差的计算有两种方法。第一种是分步进行，先用函数"MAX"与"MIN"求出每个班分数的最大值与最小值，然后作差求出。第二种是综合计算，如果比较熟悉后，可以直接列出算式计算，其计算步骤与计算平均数等类似。如图3－24，在单元格E5中键入"＝MAX (C2:C39)－MIN(C2:C39)"，按Enter键，返回值26就显示在单元格E5中。E5表示的是高二(1)班38名学生数学单元测验成绩的

极差。在单元格 F5 中键入“＝MAX(C40:C97)－MIN(C40:C97)”，按 Enter 键，返回值 76 就显示在单元格 F5 中。F5 表示的是高二(2)班 58 名学生数学单元测验成绩的极差。

5．计算标准差与方差

标准差的计算采用函数“STDEV”，其计算步骤与计算平均数、中位数类似，计算过程略。方差等于标准差的平方，也很容易得出。

6．计算偏度与峰度

偏度系数的计算采用函数“SKEW”，峰度系数的计算采用函数“KURT”，其计算步骤与计算平均数、中位数类似，计算过程略。

所有结果如图 3－24 所示。

F9 =KURT(C40:C97)

	A	B	C	D	E	F
1	班级	学号	成绩		高二(1)	高二(2)
2	1	1	98	平均分	87.39474	73.87931
3	1	2	94	中位数	89.5	76
4	1	3	91	众数	99	76
5	1	4	99	极差	26	57
6	1	5	87	标准差	8.228295	12.41685
7	1	6	77	方差	67.70484	154.1782
8	1	7	99	偏度系数	-0.28523	-0.61355
9	1	8	73	峰度系数	-1.1[illegible]2	0.232263
10	1	9	82			

例3-7 / 例3-8 / Sheet2 / Sheet3

图 3－21

解法 2 用 SPSS 软件，共分三步完成。

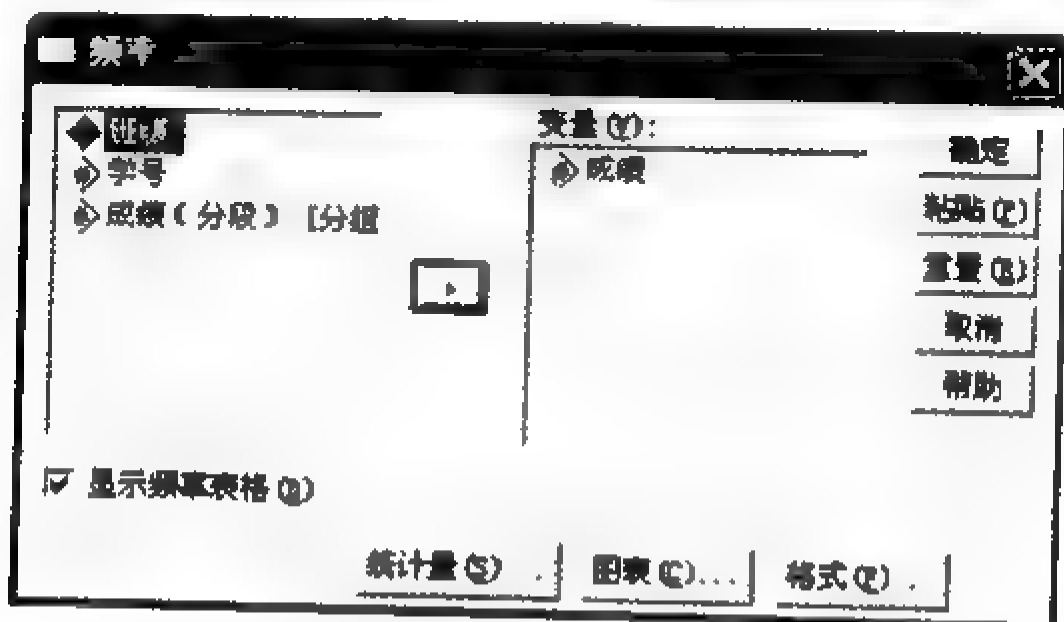
第一步，将两个班数学单元测验成绩输入 SPSS 数据编辑器的工作表区内。

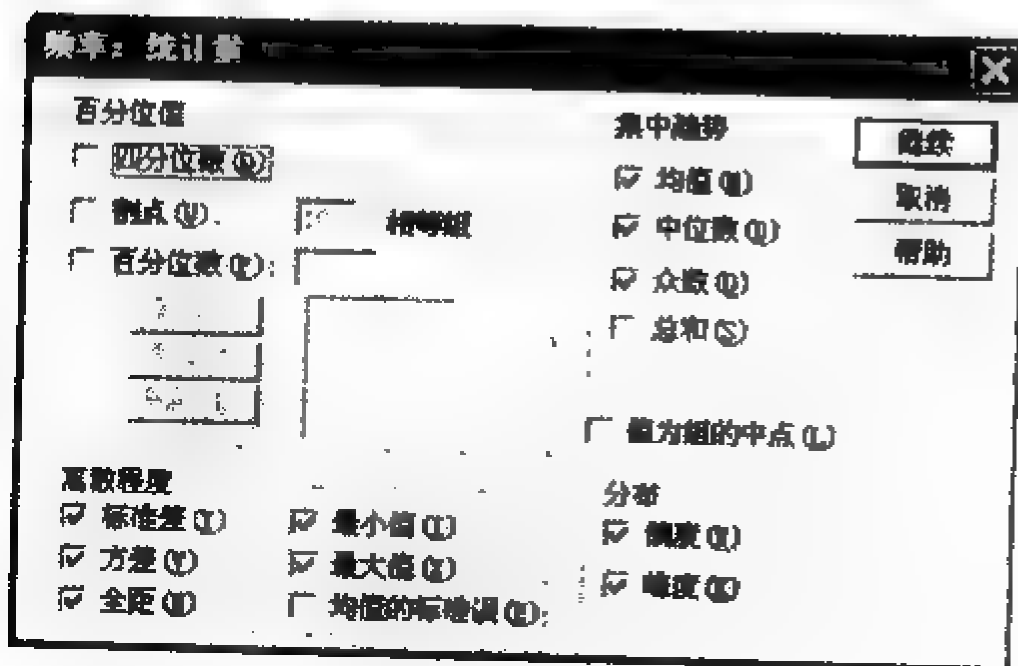
第二步，计算各种描述性统计量。

如图 3－25，执行【分析】/【描述统计】/【频率】程序，出现“频率”对话框，如图 3－26。

把左边“成绩”变量导入到右边“变量”下的空框中。然后，单击“统计量”按钮，出现“频率：统计量”子对话框，如图 3－27。


图 3－25


图 3－26


图 3－27

在图 3－27 中，“集中趋势”栏目选中“均值”、“中位数”、“众数”、“离散程度”栏目选中“标准差”、“方差”、“全距”、“最小值”、“最大值”，在“分布”栏目中选中“偏度”、“峰度”。单击“继续”按钮，回到“频率”对话框。

单击“确定”按钮，软件自动计算，然后给出结果，如表 3－11。

表 3－11　高二年级 1、2 班数学成绩统计合表

总个数	96	偏度的标准误	.246
均值	79.23	峰度	.439
中值	79.50	峰度的标准误	.488
众数	76	全距	62
标准差	12.768	极小值	37
方差	163.021	极大值	99
偏度	－.713		

统计表 3－11 中各个统计量数值，也可以执行【分析】/【描述统计】/【描述】程序，接下来的具体步骤与【分析】/【描述统计】/【频率】的类似，有兴趣的读者可以自行探索。

如果希望得到两个班各自的描述统计量的值，计算方法如下。

1. 如图 3－28，执行【分析】/【比较均值】/【均值】程序，出现“均值”对话框，如图 3－29。

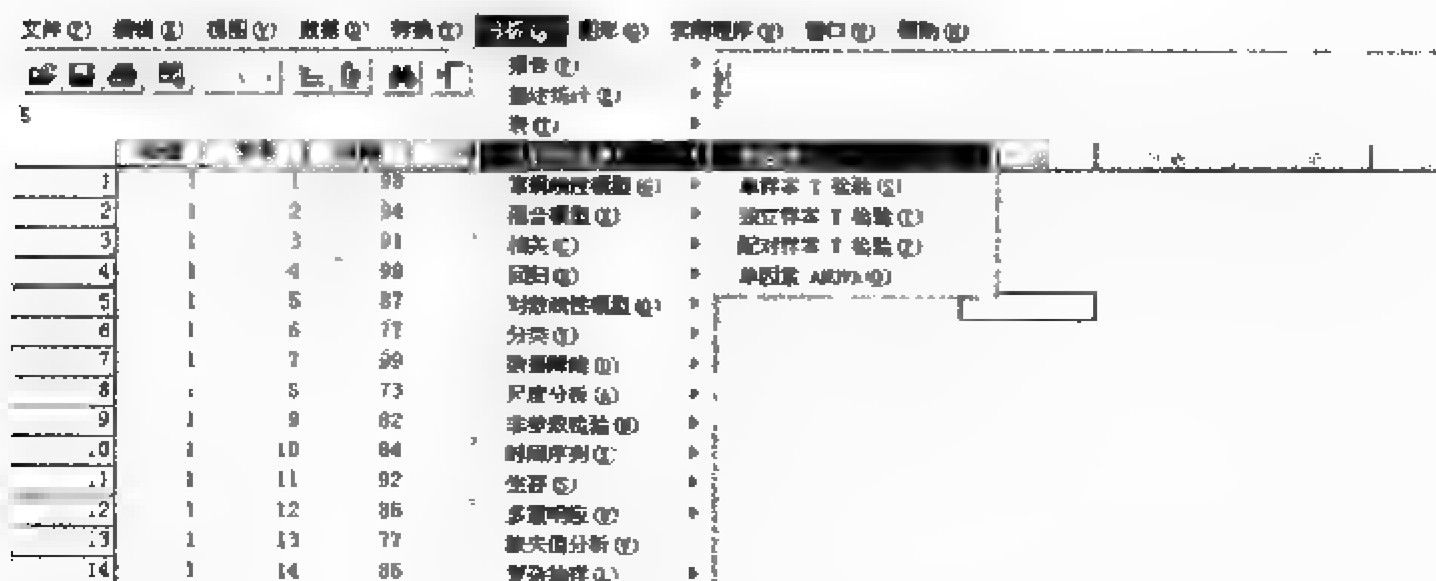


图 3　28

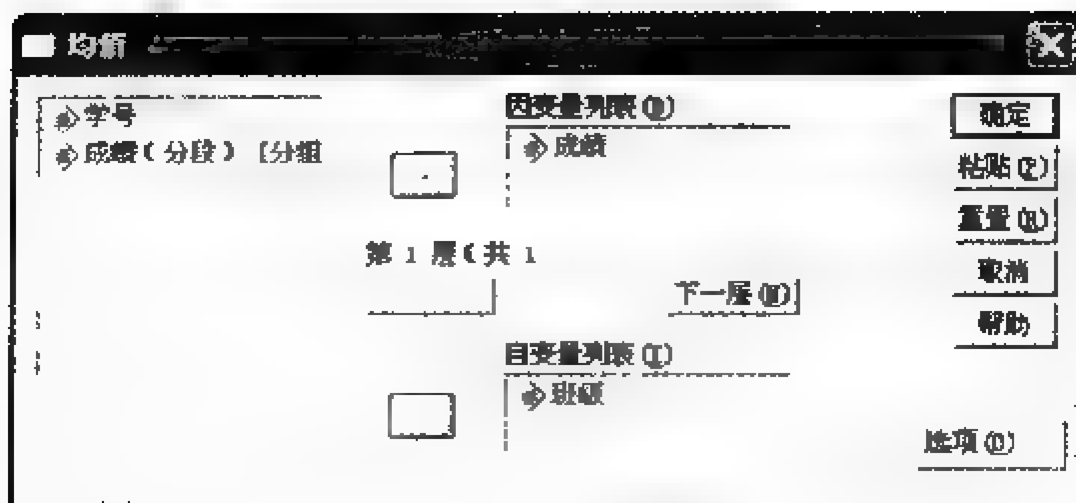


图 3－29

2. 把左边的变量“成绩”导入到右边“因变量列表”下的空框中，“班级”导入到右边“自变量列表”下的空框中。单击“选项”按钮，出现“均值:选项”子对话框，如图 3－30。

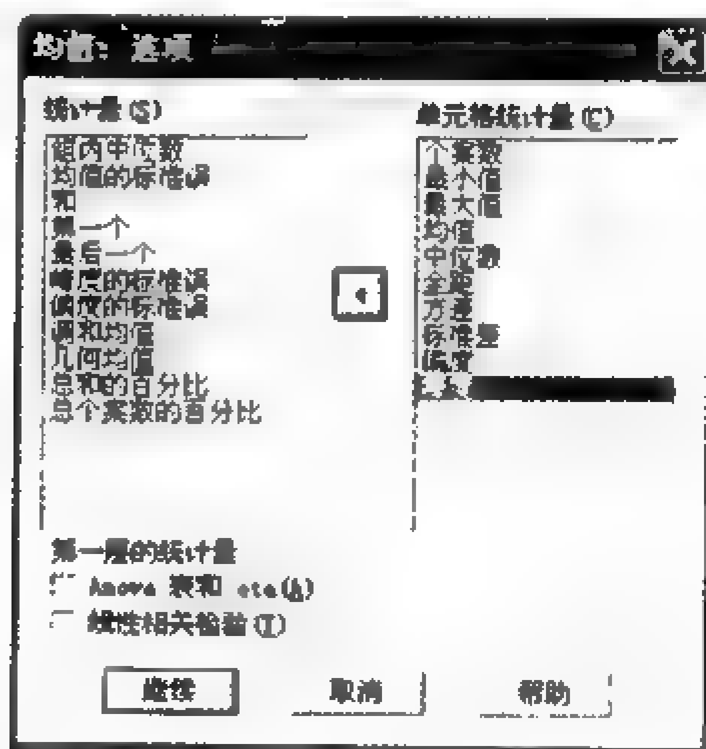


图 3－30

3. 将左边“统计量”栏目下的“个案数”、“最小值”、“最大值”、“均值”、“中位数”、“全距”、“方差”、“标准差”、“偏度”、“峰度”等统计量导入到右边“单元格统计量”下的空框中。

4. 单击“继续”按钮，返回到“均值”对话框（图 3－29）。再单击“确定”按钮，软件自动运行计算程序，给出计算结果。如表 3－12。

除了表 2－12 中列出的数值外，SPSS 软件还可以计算其他的描

述统计量，读者可以根据自己的需要自行选择。

表 2 12 高二年级 1、2 班数学单元测验成绩报告表

班级	人数	极小值	极大值	均值	中位数	全距	方差	标准差	偏度	峰度
1	38	73	99	87.39	89.50	26	67.705	8.228	−.285	−1.116
2	58	37	94	73.88	76.00	57	154.178	12.417	−.614	.232
总计	96	37	99	79.23	79.50	62	163.021	12.768	−.713	.439

三、制作统计图表

EXCEL 和 SPSS 软件可以制作各式各样的统计图表，下面以直方图、茎叶图为例，介绍作图的基本程序，其他图形可以参照此方法进行探索。

【例 3－8】 在例 3－6 的基础上，用直方图表示两个班测验成绩的分数分布。

解法 1 用 EXCEL 软件，共分三步完成。

第一步，根据分组情况，确定每个组的组中值，并输入 EXCEL 工作簿的工作表区内。考虑到测验分数是连续数据，故组中值分别取 35、45、55、65、75、85、95，如图 3－31。

	A	B	C	D	E	F	G	H	I
1	高二年级1、2班								
2	班级	学号	成绩	最值	分组	分点	1班频数	2班频数	组中值
3	1	1	98	99	[30, 39]	39	0	1	35
4	1	2	94	37	[40, 49]	49	0	1	45
5	1	3	91		[50, 59]	59	0	6	55
6	1	4	99		[60, 69]	69	0	12	65
7	1	5	87		[70, 79]	79	9	19	75
8	1	6	77		[80, 89]	89	10	12	85
9	1	7	99		[90, 99]		19	7	95
10	1	8	73						

例3-7 例3-6 Sheet2 Sheet3

图 3－31

第二步，根据频数与组中值，制作条形图。

(1) 单击“图表向导”按钮，弹出“图表类型”对话框，如图3-32。选择柱形图，单击“下一步”按钮，进入“图表数据源”对话框，如图3-33。

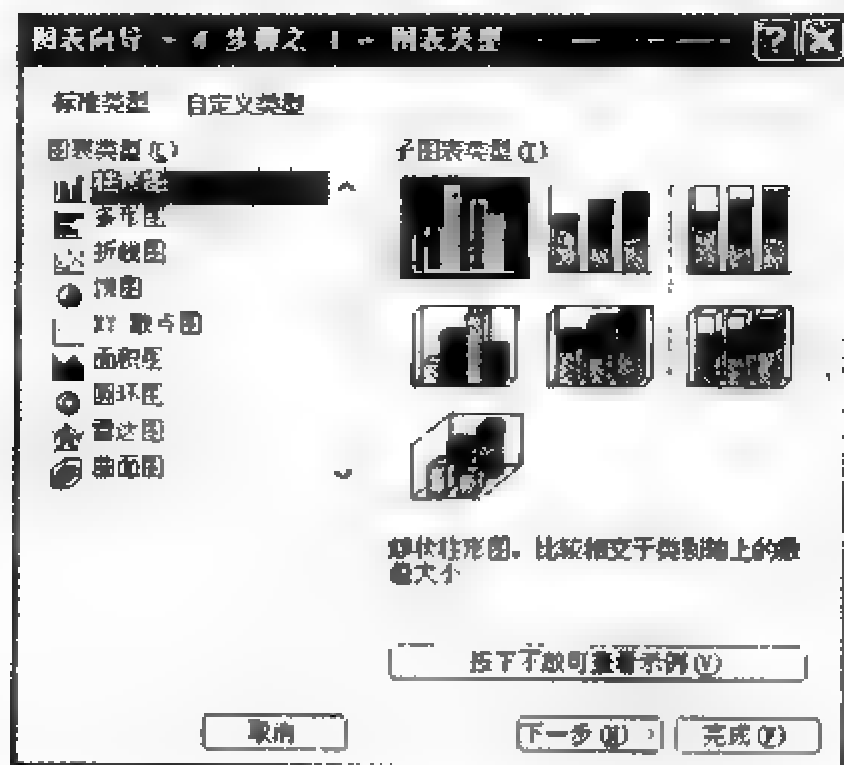


图 3-32

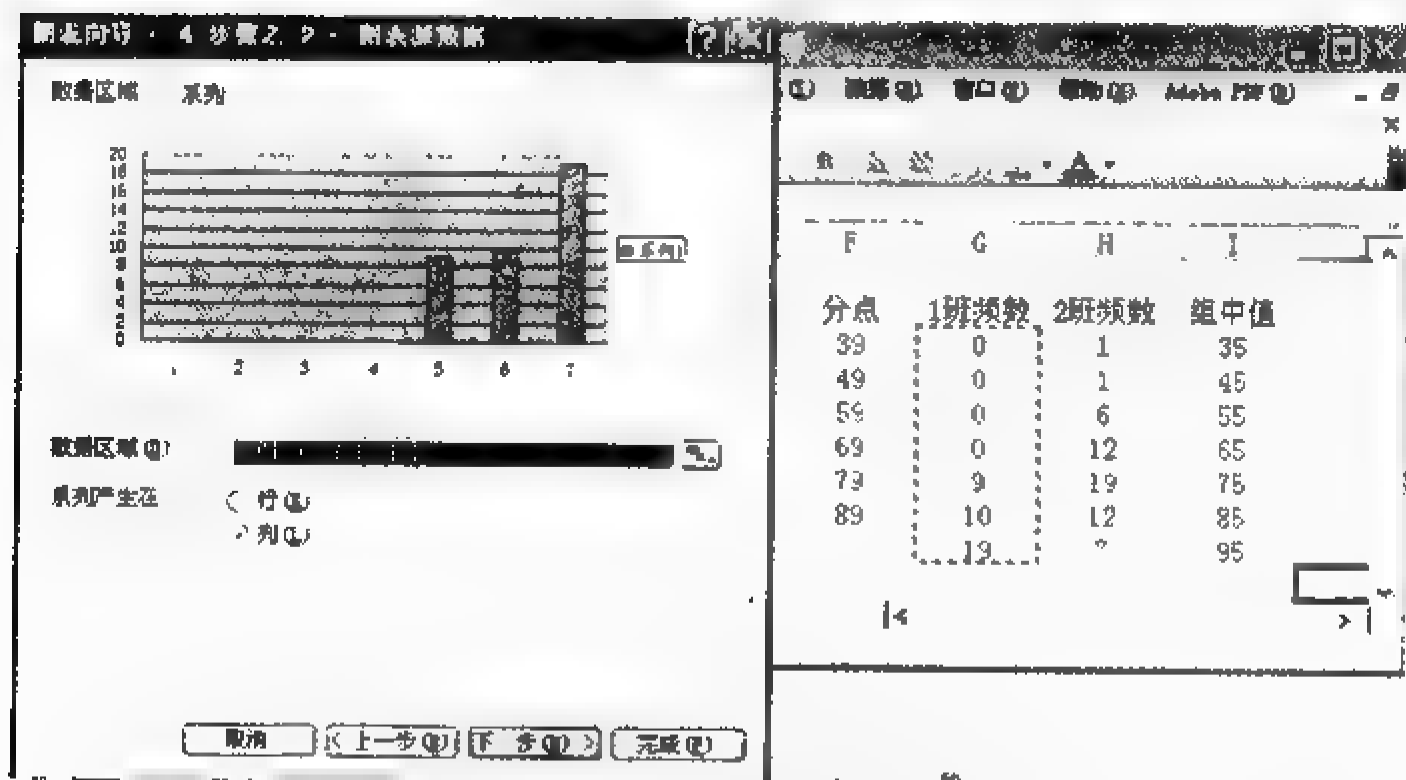


图 3-33

(2) 在“数据区域”选项卡中，单击“数据区域”选项右边的图标，出现工作表，用鼠标选中“1班频数”下面的数据；则“数据区域”

选项卡下方呈现条形统计图。

单击“系列”选项卡，在“系列”选项卡中，单击“分类(x)轴标志”选项右边的图标[图标]，出现工作表，用鼠标选中“组中值”下面的数据；则“系列”选项卡下方条形统计图的横轴出现用组中值 35、45、55、65、75、85、95 作为长条的标识，如图 3－34。

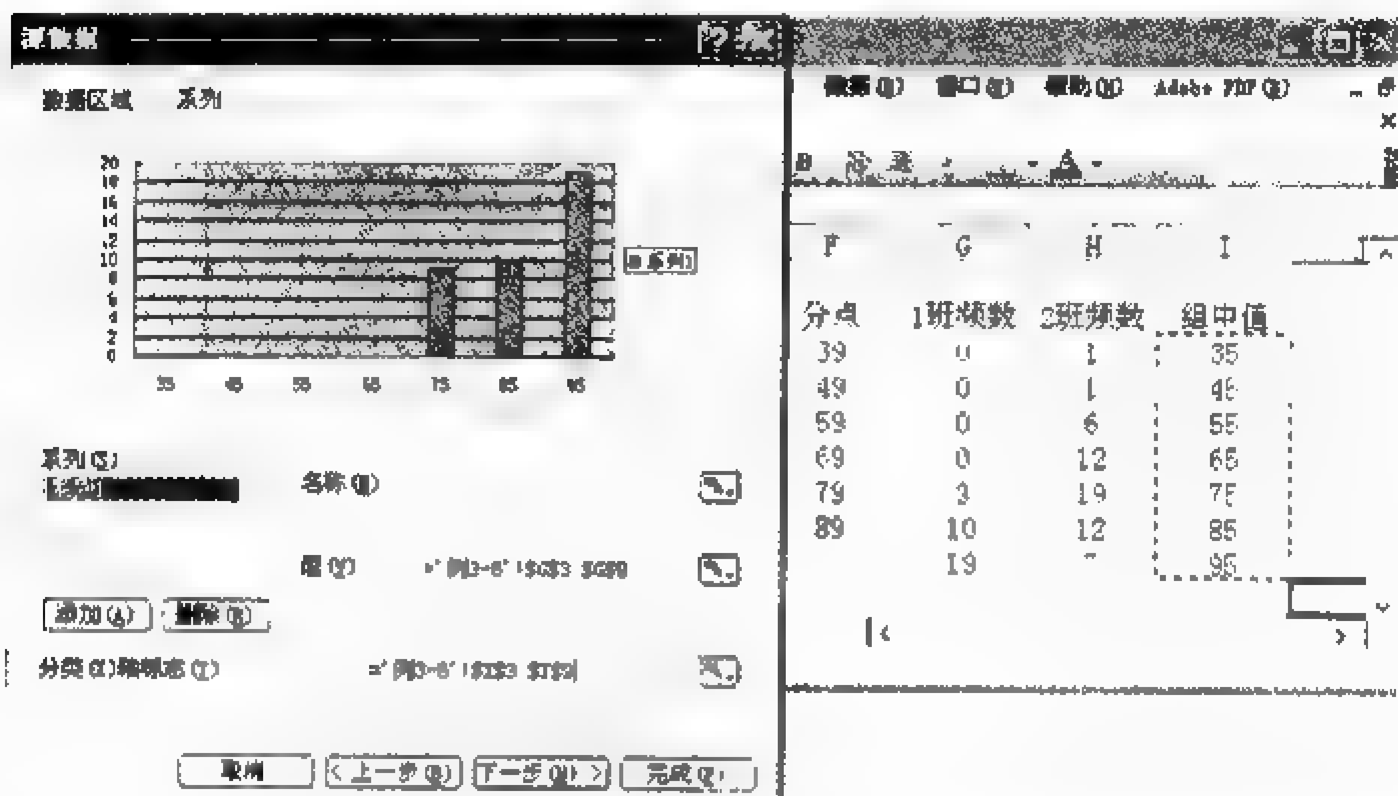


图 3－34

(3) 设计好数据引用后，单击“下一步”按钮，进入“图表选项”对话框，如图 3－35。选定“标题”选项卡，在“图标标题”栏目下输入

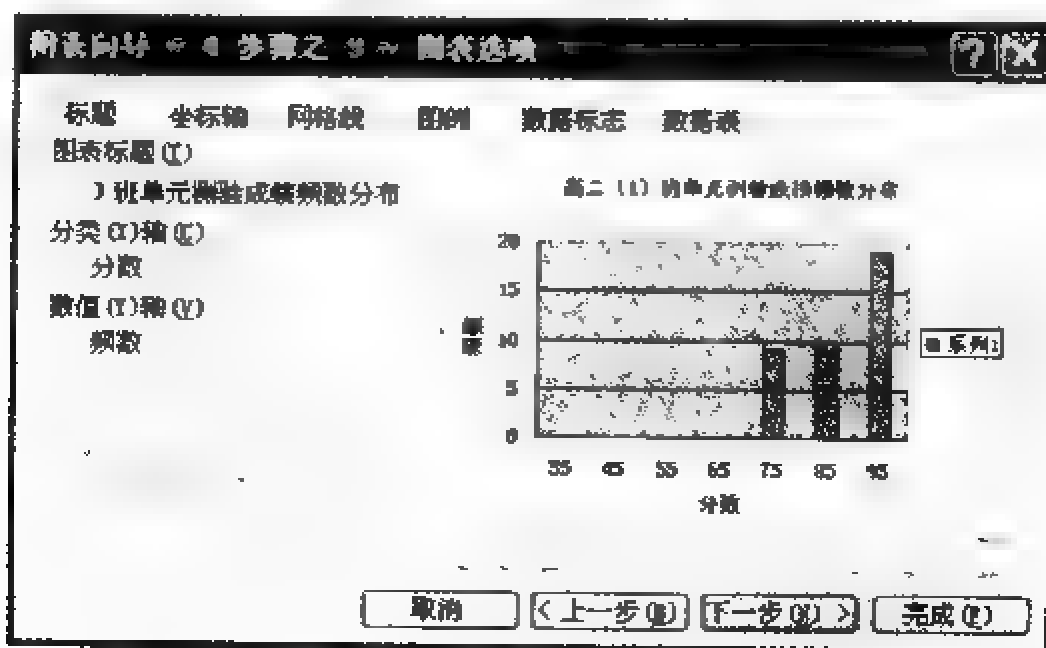


图 3－35

“高二(1)班单元测验成绩频数分布”，在“分类(X)轴”栏目下输入“分数”，在“数值(Y)轴”栏目下输入“频数”。依次选择其他选项卡“坐标轴”、“网格线”、“图例”、“数据标志”、“数据表”，根据需要进行设置。

(4) 设置好图表的各个选项后，单击“下一步”按钮，进入“图表位置”对话框，如图3-36。若选择将图表“作为其中的对象插入”，单击“完成”按钮，则最后生成图3-37所示的图表。

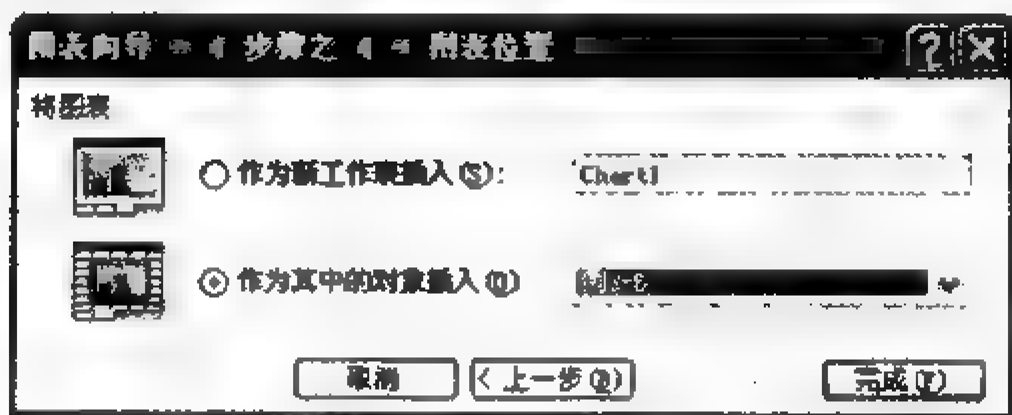


图 3-36

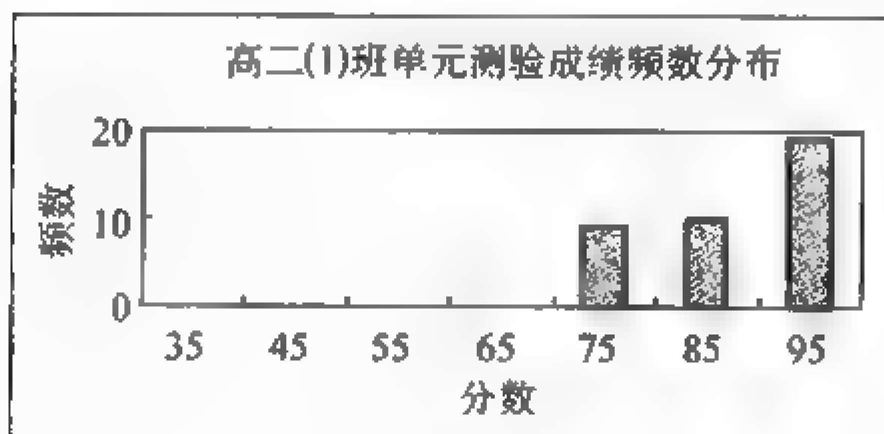


图 3-37

第三步，修改条形图，得到直方图。

1. 如图3-38，在EXCEL工作表中，用鼠标在任一直长条上单击，选中长条。再右击鼠标出现下拉菜单。选中“数据系列格式”后，出现“数据系列格式”对话框，如图3-39。

2. 在“选项”卡中将“分类间距”设置为零，再按“确定”按钮。则图3-37就修改成图3-40，即得到直方图。

至于高二(2)班单元测验成绩的频数分布直方图的制作方法，与上述步骤一样，其过程略。图形见图3-3。

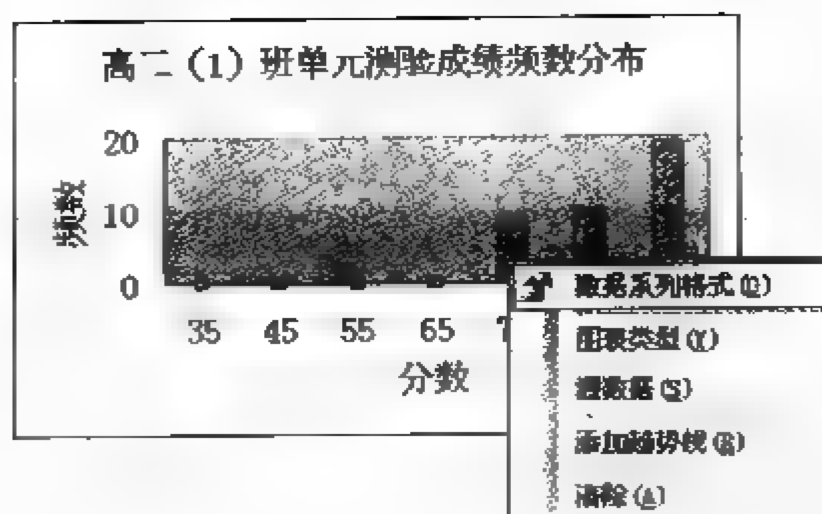


图 3－38

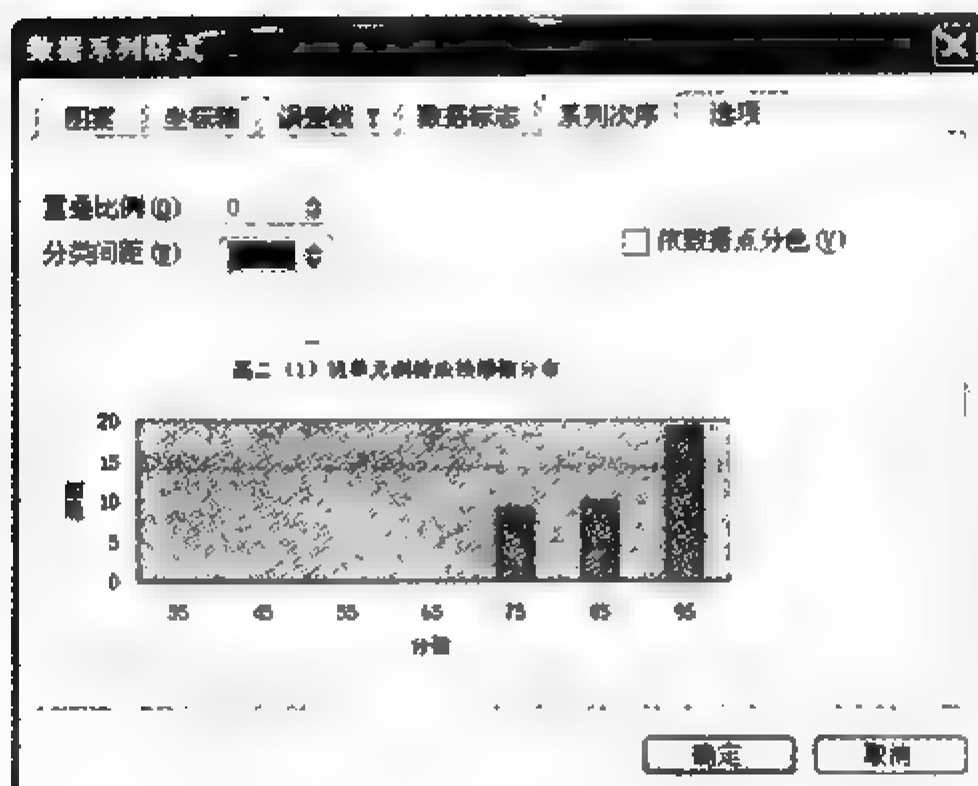


图 3－39

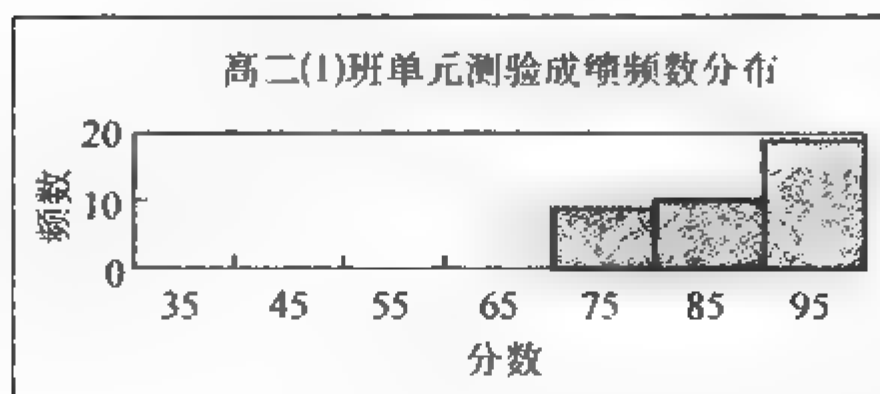


图 3－40

解法 2 用 SPSS 软件制作直方图比较简单。

首先制作两个班总测验成绩的频数分布直方图。

1. 如图3－41，首先，打开例3－6文件，执行【图形】/【直方图】程序，出现“直方图”对话框。

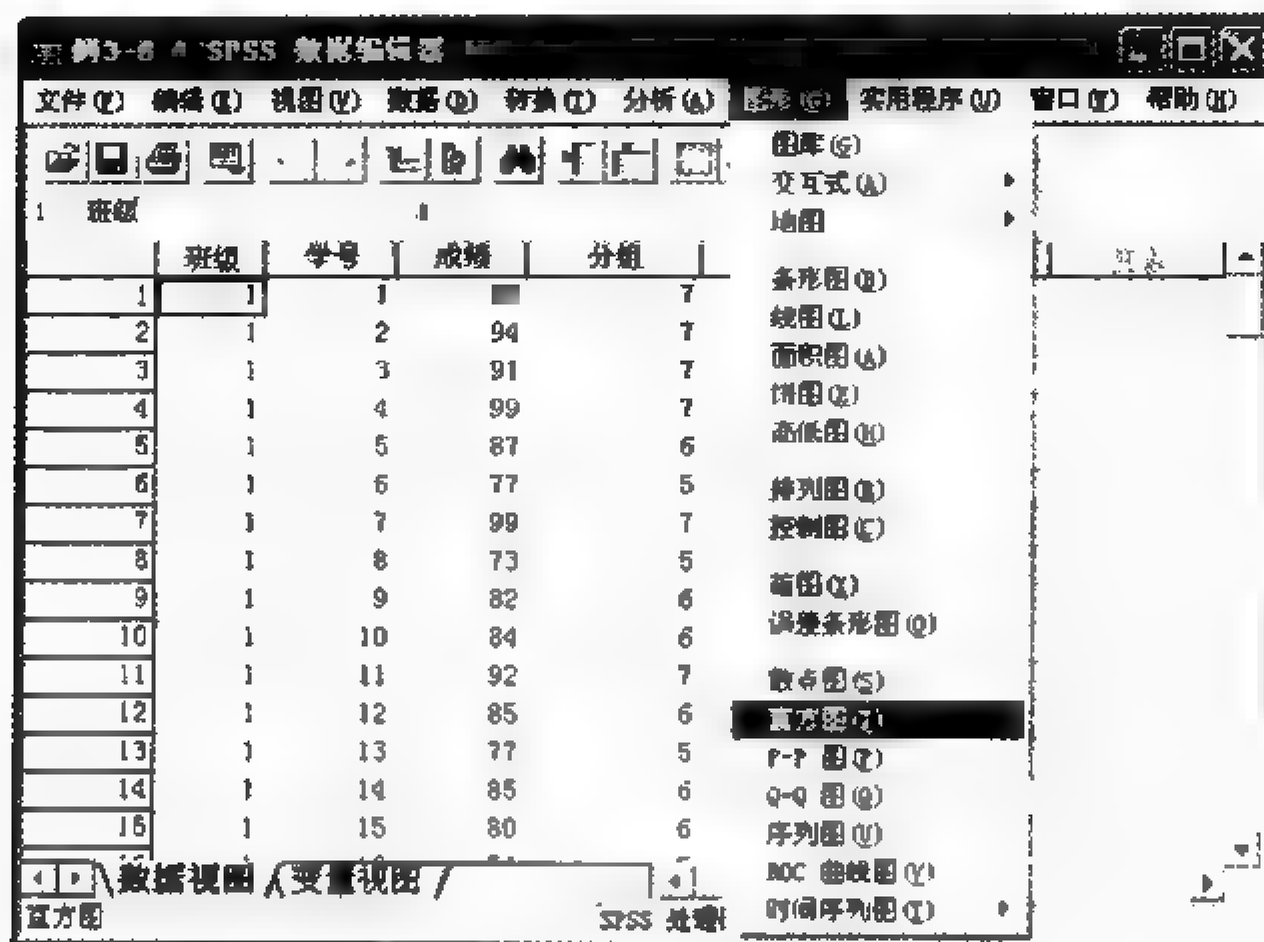


图 3－41

2. 如图3－42，将左边“成绩”变量导入到右边“变量”栏目下的空框中，单击“确定”按钮，得到图3－43。

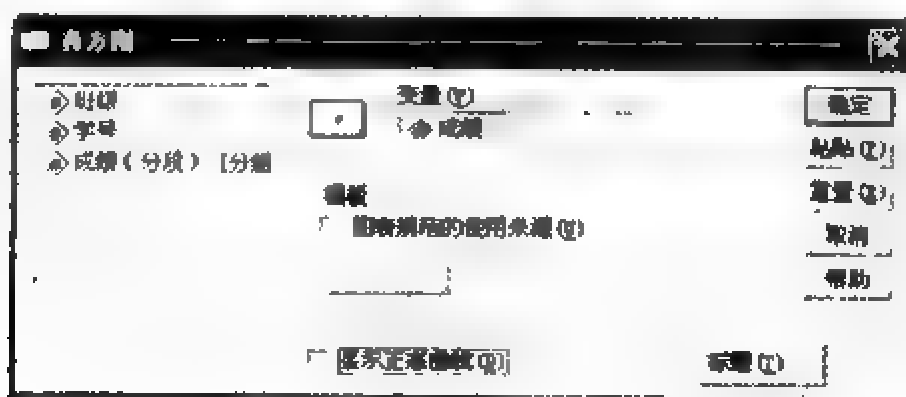


图 3－42

图3－43表示的是高二(1)、高二(2)两个班整体的频数分布直方图，图中组距是5，共分为13个组。

下面介绍同时制作高二(1)班、高二(2)班两个班测验成绩的频数分布的直方图的方法。

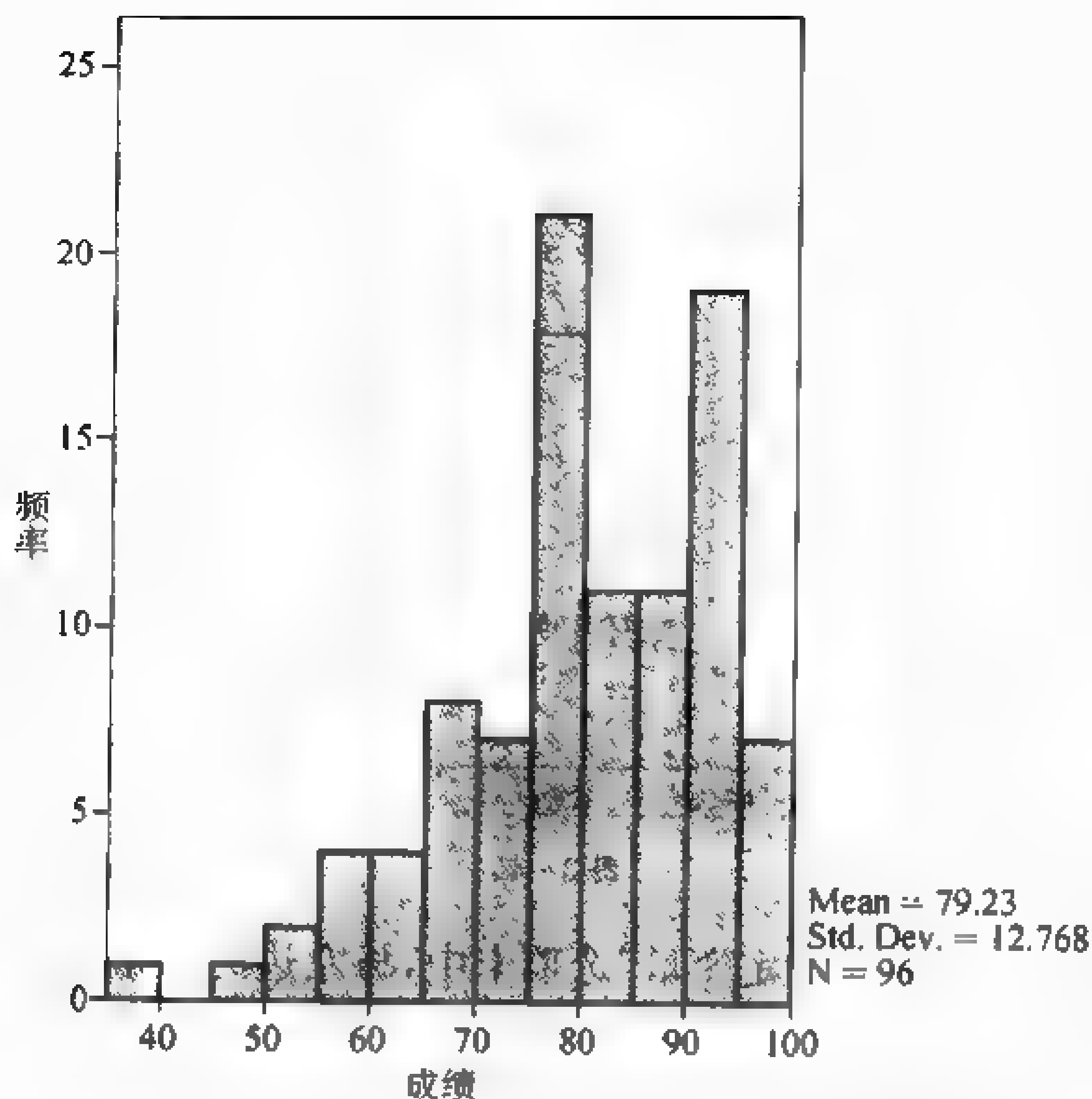


图 3-43　高二(1)、高二(2)两个班整体的频数分布直方图

1. 如图 3-44，可以执行【图形】/【交互式】/【直方图】程序，得到"创建直方图"对话框，如图 3-45。

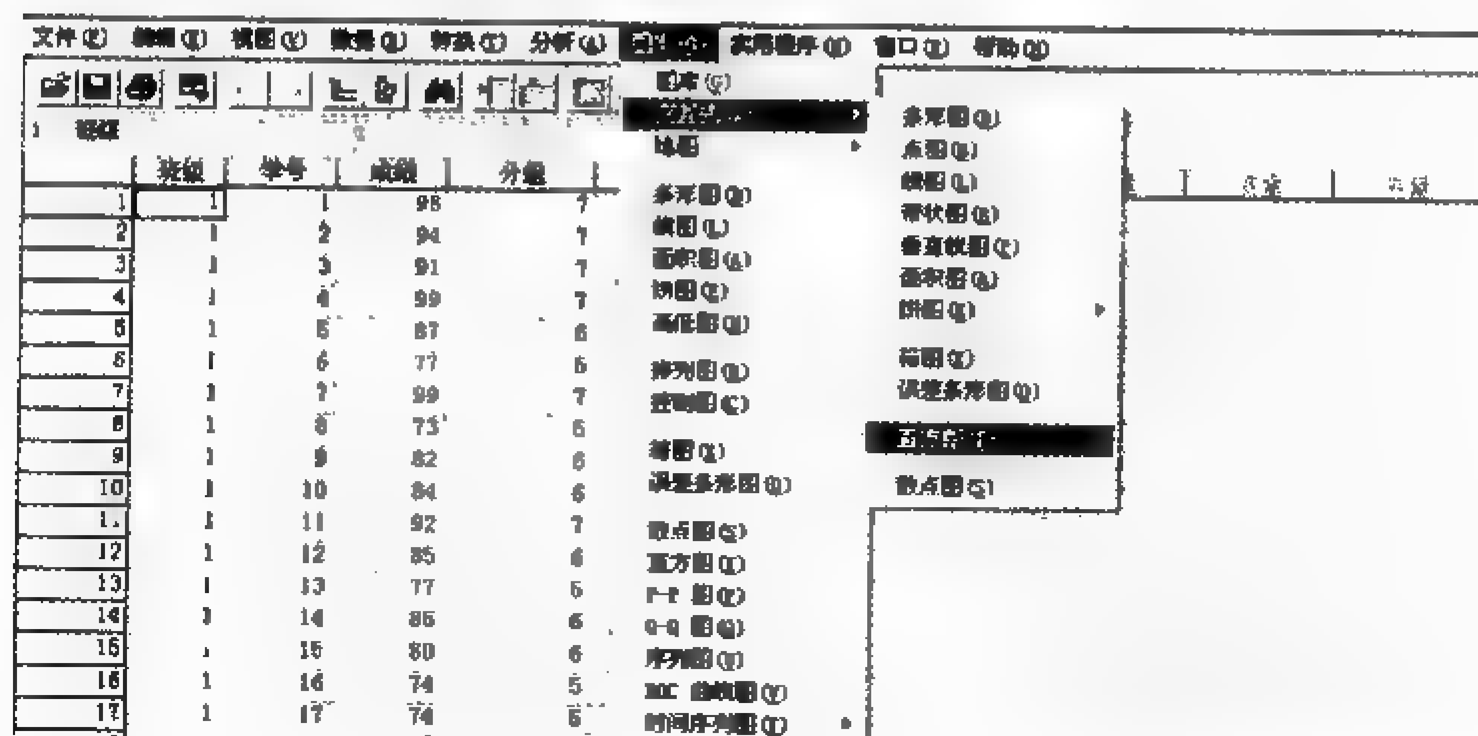


图 3-44

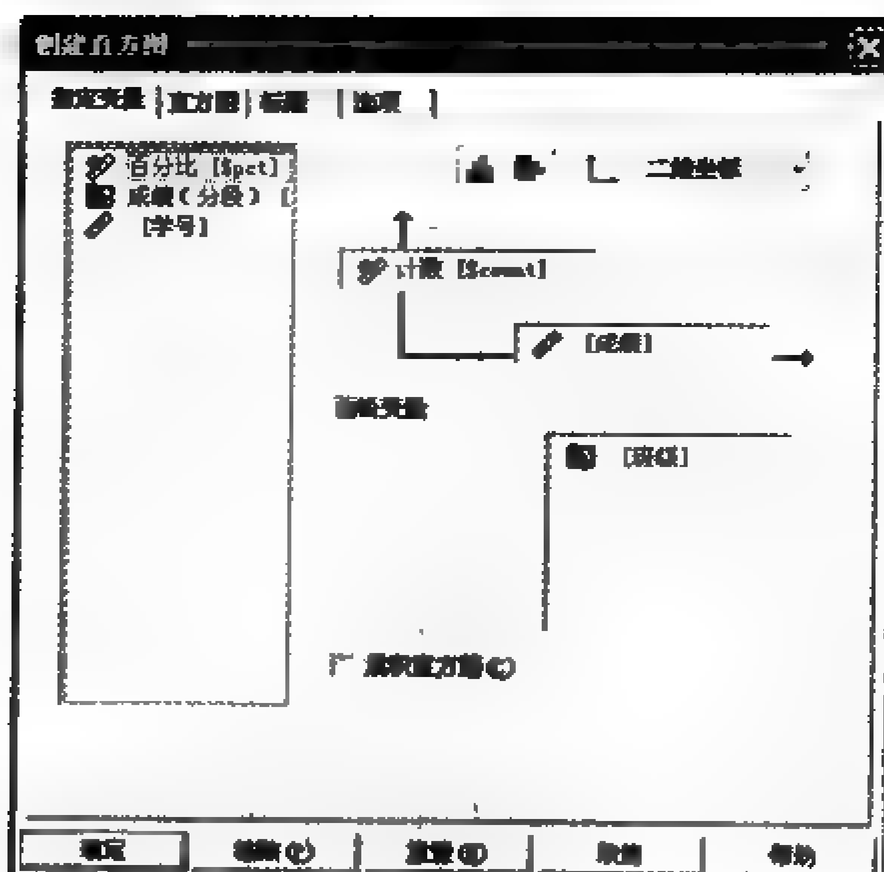


图 3－45

如图 3－45，在“指定变量”选项卡中，将左边的变量“班级”导入到右边“面板变量”下的空框中；将左边的变量“成绩”导入右边 x 轴相应的空框中。

如图 3－46，在“直方图”选项卡中，设置“区间大小”栏目，将“区

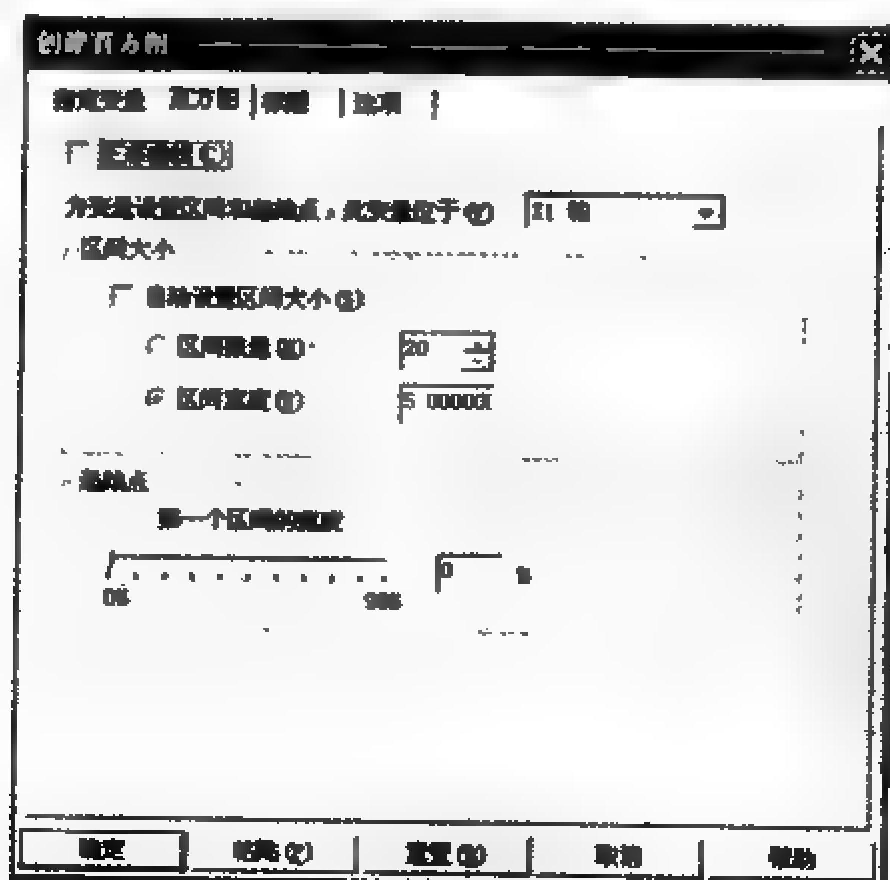


图 3－46

间数量”设置为 20，将“区间宽度”设置为 5。还可以根据需要设置其他选项卡。设置完毕后，单击“确定”按钮。软件运行后，得到图 3－47，它表示的是两个班成绩的频数分布直方图的对比。需要注意的是，这里的分组，高二(1)班、高二(2)班的数据第一组区间分别是［73，78)、［37，42)，依次类推。因此，SPSS 软件给出的频数分布直方图与 EXCEL 软件给出的有差异，这一点，请读者注意。

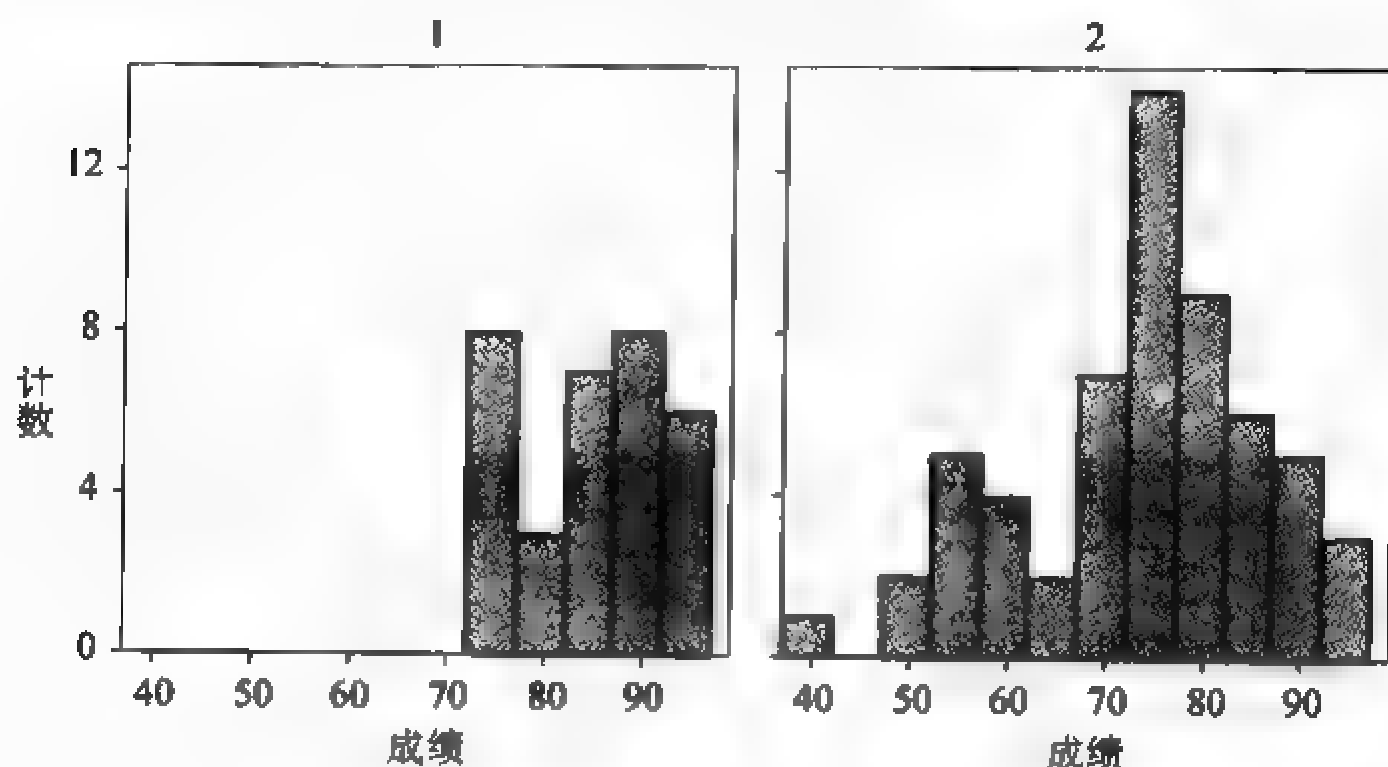


图 3－47

【例 3－9】 用 SPSS 软件制作例 3－6 中两个班测验成绩分数分布的总茎叶图。

解：茎叶图的制作分 3 步进行。

1. 如图 3－48，执行【分析】/【描述性统计】/【探索】程序，出现“探索”对话框，如图 3－49。

2. 在图 3－49 中，将左边的“成绩”变量导入到右边“因变量列表”栏目下的空框中。单击“图”选项，出现“探索：图”子对话框，如图 3－50，在“描述性”栏目下选中“茎叶图”，然后单击“继续”按钮，返回到图 3－49。

3. 单击“确定”按钮，即可得到茎叶图。如图 3－51 所示。

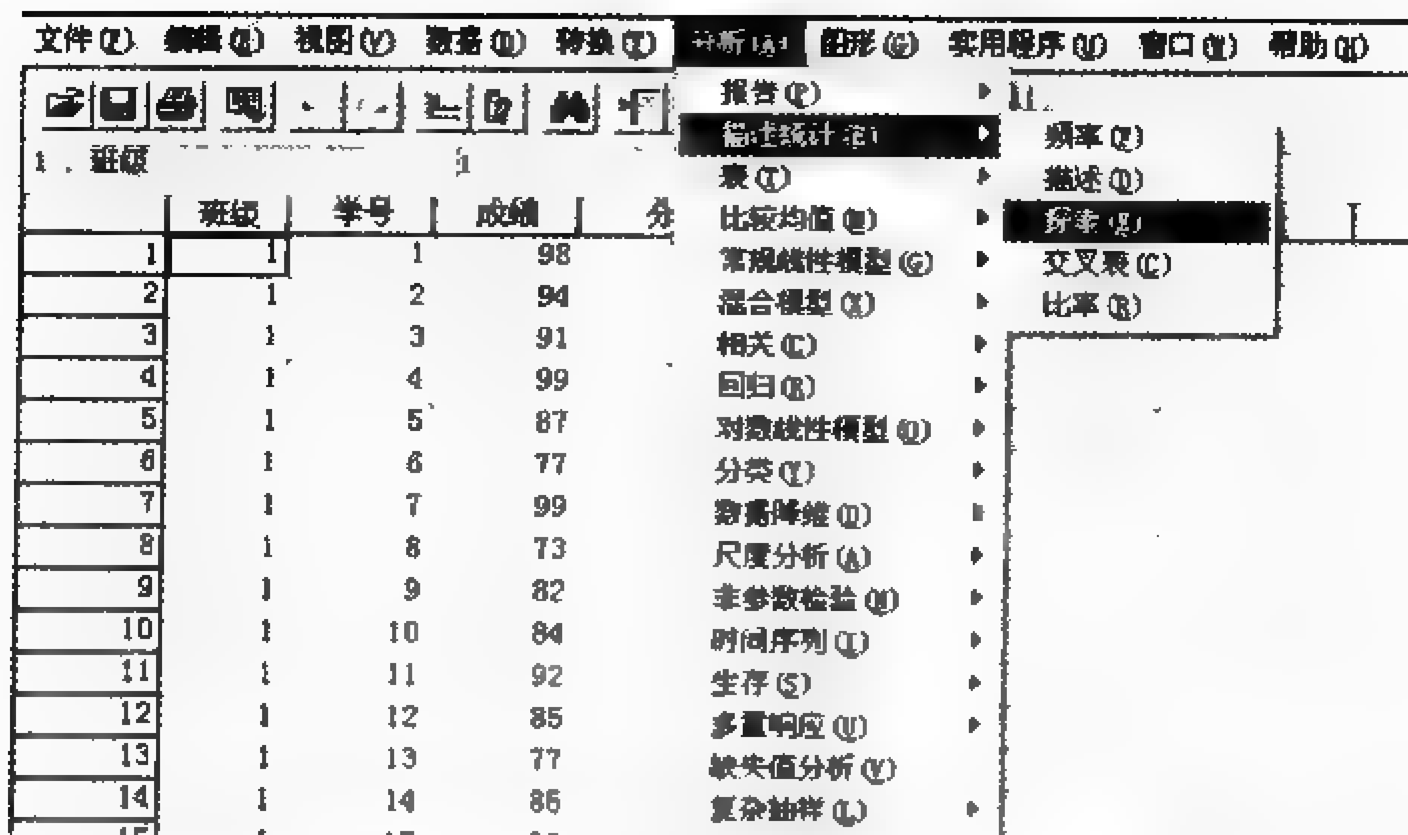


图 3－48

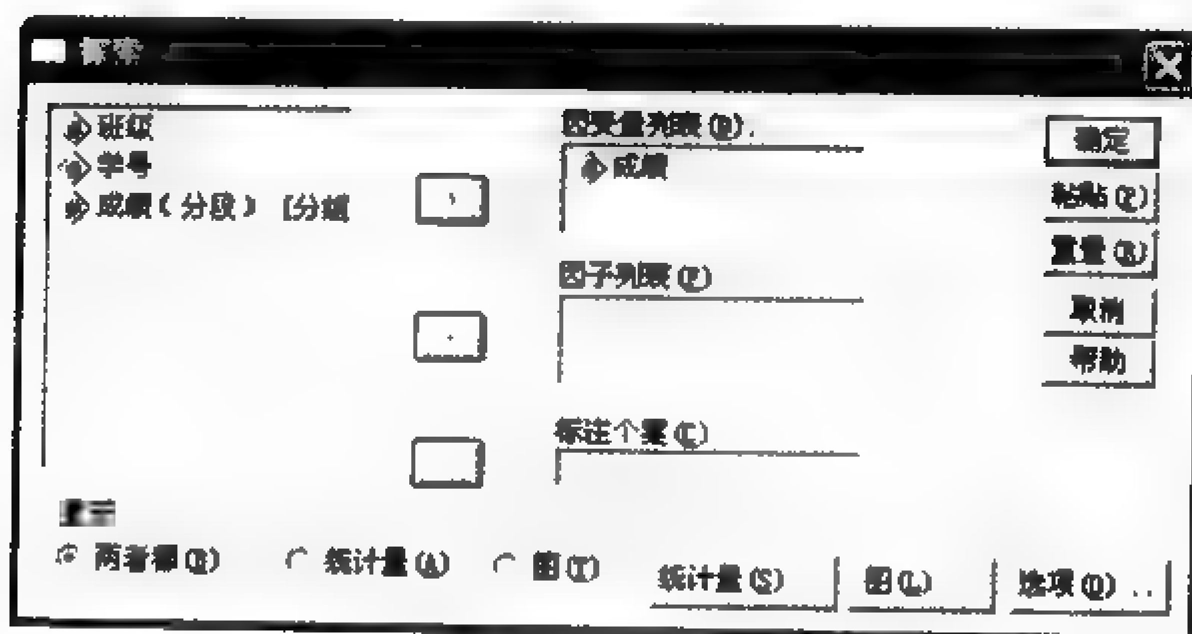


图 3－49

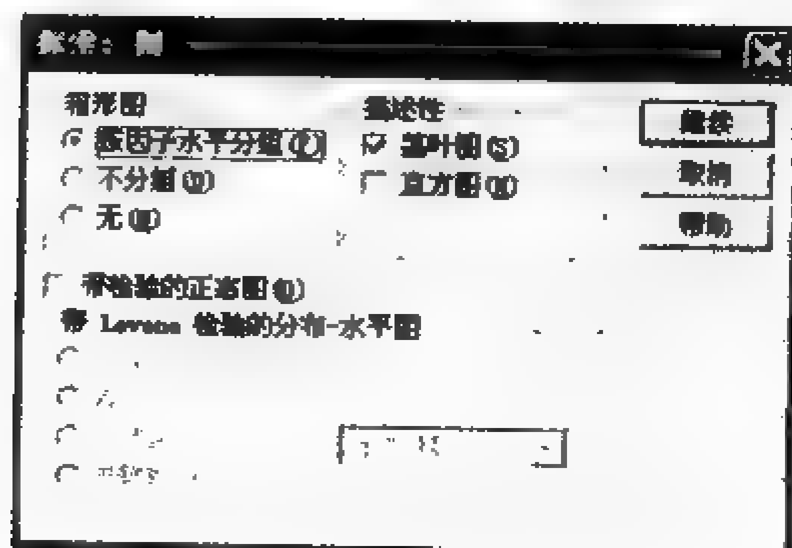


图 3－50

```
成绩 Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     2.00 Extremes    (=<48)
     2.00        5 .  24
     4.00        5 .  5577
     4.00        6 .  0022
     8.00        6 .  56888899
     7.00        7 .  2334444
➡   21.00        7 .  555556666666777788899
    11.00        8 .  00001124444
    11.00        8 .  55555777799
    19.00        9 .  0000111112233334444
     7.00        9 .  5889999

 Stem width:        10
 Each leaf:       1 case(s)
```

图 3-51

将图 3-51 翻译成中文，如下表 3-13。

表 3-13　两个班单元测验成绩的总茎叶图

频数	茎	叶
2.00	极端值	(=<48)
2.00	5.	24
4.00	5.	5577
4.00	6.	0022
8.00	6.	56888899
7.00	7.	2334444
21.00	7.	555556666666777788899
11.00	8.	00001124444
11.00	8.	55555777799
19.00	9.	0000111112233334444
7.00	9.	5889999
茎宽：	10	
每叶代表数据数：	1 个数据	

第四章 测验成绩的统计分析

运用推断统计对整理后的测验成绩进行较深层次的比较、研究，主要包括测验质量指标的计算与分析，考生不同测验成绩之间的分析与归因，班级之间、班级与年级之间分数的比较与分析，测验成绩的影响因素分析，根据已有成绩预测重要考试的结果等，这些研究内容使用的基本技术与方法包括相关分析、回归分析与方差分析。

第一节 相关分析

在教育测验中，测验信度、效度、试题区分度等质量指标都是经常研究的内容。测验信度是指平行测验中任意两个测验间测验结果的一致性程度，通常用两个测验结果的相关系数来表示。测验的效标关联效度是用测验分数和效标之间的相关系数表示测验效度的高低，效标就是检测效度的参照标准。试题区分度是指测验试题鉴别考生实际能力水平高低的量度，实际操作时，把考生在整个测验上所得的总分当成考生的实际能力水平，其中一种方法就是用考生群体在试题上的得分与测验总分的相关系数来表示试题的区分度。除了这些测验质量指标外，我们还需要研究试题间的关系、不同测验间的关系等，这些问题的解决都需要用到相关性与相关分析。

一、相关与相关系数

1. 相关与相关分析

在教育研究实践中，很多研究对象之间存在相互关系但不能做出因果解释。当事物间存在联系又不能直接做出因果关系解释时，称事物之间的联系为相关。例如，一份数学单元测验卷由客观题和主观题两部分组成，考生的客观题得分与主观题得分之间存在一定的联系，因为都反映出考生对特定知识技能的掌握情况，但同一考生两部分得分之间并不存在"因"与"果"的关系。

用一些合理的指标对相关的事物的观测值进行统计分析的方法称为相关分析。

2. 相关系数

衡量两个或多个变量间相关程度的定量化指标称为相关系数，用 r 表示，r 具有以下五个基本特点。

(1) 相关系数是同一个考生群体参与测验得出的两组实测分数间的数量指标，或者是基于相同基础成对匹配的两个考生群体同一次测验实测分数间的数量指标。

(2) r 的取值范围为 $-1 \leqslant r \leqslant 1$，$r$ 值的符号表示变量之间关系的方向。当 $r > 0$ 时，变量之间正相关；特别地，当 $r = 1$ 时，变量之间严格正相关。当 $r < 0$ 时，变量之间负相关；特别地，当 $r = -1$ 时，变量之间严格负相关。$r = 0$ 表示变量之间不存在相关关系。

(3) r 值越大，表明变量之间的关系越强。

(4) r 仅仅表示变量间的联系程度，即使 $|r|$ 值很大，也不能说明变量之间存在因果关系。

(5) 相关系数间只是存在大小关系，不存在倍数关系。例如 $r_1 = 0.8$ 的关联程度仅仅表示比 $r_2 = 0.4$ 的关联程度大，而不能说前者变量间关联程度是后者变量间关联程度的 2 倍。

3. 相关系数的解释

寻找两个变量之间的相关关系时，人们通常借助散点图展开研

究，探索变量间是否相互联系，可能以什么方式相互联系；然后计算相关系数，进一步探索变量之间怎样的联系才是好的相关？相关系数值达到多少时，才能说变量间的相关关系显著？判断相关系数显著性的依据、方法与标准是什么？这些问题都是使用相关系数时需要解决的问题。

(1) 相关系数的重要性依赖于研究目的

“变量之间怎样的联系才是好的相关”是如何认识相关系数重要性的问题。相关系数的重要性依赖于研究的种类，研究希望解决的问题。例如，在一个陌生领域探索两个变量间的联系时，相关系数值 0.15 也许极其重要，然而在教育测验中用于判定某道试题与测验总分的联系时，这个相关系数值很不理想，这道试题应予淘汰。

(2) 解释相关系数时需要考虑数据的背景

在解释相关系数时，尤其需要考虑数据收集的环境和基于统计希望做出什么决定。如果两个变量相关，那么这两个变量之间的关系可能有多种。对于教育测验而言，可以结合以下几种关系来解释相关系数。

① 两个变量中，可能一个变量确定了另一个变量，即可能两者之间含有因果关系。例如，测验总分与某道试题的得分关系，如果假定测验总分代表考生能力水平，那么能力水平高的考生在解答题难度较高的题目上得分相应也高，即难度较高的题目与测验总卷的相关系数值往往较高。

② 两个变量可能是某一共同因素的相关效应。例如，在数学测验中，试题与试题之间的相关性往往受到考生数学能力的共同影响。

③ 两个变量测量了某种共同的能力。例如，在中考、高考等大规模的考试中，一份数学测验卷包含的知识点很多，那么有些试题可以归为一类，测量的是考生数学中某一部分的能力。

(3) 解释相关系数时需要考虑使用的数据情况

在解释相关系数时，还需要考虑收集的数据具有的特征。由于

相关系数的计算与所收集的数据的样本容量、样本平均数等都有密切的关系，因此，在解释相关系数时，需要说明计算涉及的数据背景、结果的适用范围。

① 样本容量问题。在解释相关系数时，需要确定基于所选定的样本，其样本容量的大小足够代表需要描述的总体，这样基于相关系数值所作出的解释才具有说服力。需要注意的是，并不是样本容量大，相关系数值就大；也不是样本容量小，相关系数值就小；关键是样本的选取必须具有典型性与代表性。

② 样本中极端值的问题。由于相关系数的计算需要用到平均数，而平均数受极端值影响很大，因此，在计算相关系数前，应借助散点图观察极端值的分布情况，确定是否把极端值包括在样本和研究中，以及如何解释极端值现象。

③ 多总体问题。在计算相关系数前，需要仔细研究数据对的呈现状况。如果数据点明显地呈现出两个群体，那么应该把数据分成两个子总体，分别加以详细研究，而不是笼统地说明数据组之间的相关性。

4. 决定系数与非决定系数

由于相关系数存在正、负两种符号，因此，在统计学中另一种解释相关系数的方法是把它的值加以平方。

相关系数的平方 r^2 被称为决定系数，它说明一个变量的方差有多少百分比可以用另一个变量来解释。如果两个变量间的相关系数为 $r = 0.685$，决定系数 $r^2 = 0.469$ 表示一个变量方差的大约 46.9%可以由另一个变量来解释。

相应地，也可以计算 $1 - r^2$，它被称为非决定系数，说明一个变量的方差有多少百分比不可以用另一个变量来解释，即是由其他的未确定因素导致的。

二、相关系数的类型及其计算

收集的数据类型不同，数据的分布情况不同，计算相关系数时

选择的公式也不同。表 4-1 根据测量量表适用的不同将测验数据分为三类，列出教育测验中经常遇到的几种相关系数类型。

表 4-1 不同类型的数据和它们适用的相关系数类型

		变量 2		
		定类数据	定序数据	定距或定比数据
变量 1	定类数据	二分相关 r_ϕ	四分相关 r_t 肯德尔(Kendall) 和谐系数 τ	点二列相关系数 r_{pb}
	定序数据	四分相关 r_t 肯德尔(Kendall)和谐系数 τ	斯皮尔曼(Spearman)等级相关系数 r_s	二列相关系数 r_b
	定距或定比数据	点二列相关系数 r_{pb}	二列相关系数 r_b	皮尔逊(Pearson)积差相关系数 r_{xy}

表中的点二列相关系数、二列相关系数、皮尔逊积差相关系数等在第二章第四节试题区分度的计算部分，已经给予了详细说明，这里不再一一赘述。下面分别简要说明二分相关、四分相关、肯德尔和谐系数与斯皮尔曼等级相关系数的计算。

1. 二分相关 r_ϕ

二分变量是指变量只包含两个类别。如：性别分为男与女，地区分为城市与农村。当两个变量都是二分变量时，描述这两个变量之间的相关就称为二分变量相关系数 r_ϕ，也简称为 ϕ 系数。只要其中有一个是真正的二分变量(如性别)，先整理出一个 2×2 列联表，如表 4-2，根据列联表的数值，就可以计算 ϕ 系数。

表 4-2 2×2 列联表

	y_1	y_2	Σ
x_1	a	b	$a+b$

（续表）

	y_1	y_2	Σ
x_2	c	d	$c+d$
Σ	$a+c$	$b+d$	$a+b+c+d$

ϕ 系数的计算公式如下：

$$r_\phi = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}。\quad (4.1)$$

其中，a、b、c、d 的含义分别表示表 4－2 中各个变量的统计频数。

【例 4－1】 某小学为研究性别与数学学习之间的关系，从全校随机抽取 300 名学生，以期末考试的数学成绩 85 分（满分 100 分）为界进行分类，如表 4－3，求性别与数学成绩间的相关系数。

表 4－3 300 名小学生某次期末数学成绩分布表

	85 分以上（含 85 分）	85 分以下	合计
男生	93	54	147
女生	87	66	153
合计	180	120	300

解：将表 4－3 中的数据代入公式 4.1 中，得

$$r_\phi = \frac{93 \times 66 - 54 \times 87}{\sqrt{147 \times 153 \times 180 \times 120}} = 0.065。$$

即该校性别与数学成绩间的相关系数为 0.065，说明该小学中性别的差异与数学成绩间没有显著的相关性。

2. 四分相关 r_t

当两个变量都是服从正态分布的连续变量（变量取值是定距或定比数据，如，测验成绩、身高、体重等），而且两个变量都被人为地分成二分变量（如，测验成绩分为合格与不合格，身高分为高与矮，

体重分为达标与不达标等），这时，表示这两个变量之间的相关，称为四分相关，其计算公式为

$$r_t = \cos\left(\frac{\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}\right)\pi。 \tag{4.2}$$

【例4-2】 某中学在新课改实施过程研究物理与数学学习之间的关系，现从全校初二、初三两个年级共随机抽取200名学生，以学年期末考试的成绩60分（满分100分）为界分为及格与不及格两类，如表4-4，求物理与数学成绩间的相关系数。

表4-4 200名初二、初三学生物理与数学成绩分布表

	及格（数学）	不及格（数学）	合计
及格（物理）	100	35	135
不及格（物理）	25	40	65
合　　计	125	75	200

解：将表4-4中的数据代入公式4.2中，得

$$\begin{aligned} r_t &= \cos\left(\frac{\sqrt{35\times25}}{\sqrt{100\times40}+\sqrt{35\times25}}\right)\pi \\ &\approx \cos 0.31867\pi \\ &\approx 0.54。\end{aligned}$$

即该中学学年期末物理与数学成绩的相关系数约为0.54。

3. **肯德尔（Kendall）和谐系数 τ**

当两个变量的数值以多于2个等级的顺序或类别表示时，这两个变量之间的一致性程度称为肯德尔和谐系数。其计算公式如下：

$$\tau = \frac{SS_R}{\frac{1}{12}k^2(n^3-n)}。 \tag{4.3}$$

其中 $SS_R=\sum_{i=1}^{n}(R_i-\overline{R})^2$ 表示变量1的 n 个数值的离差平方和；n 表示变量 1 的等级个数；k 表示变量 2 的类别个数。

如果要研究 6 位任课教师（$k=6$）对 5 位学生（$n=5$）数学能力水平排序评价的一致性程度，可以采用公式 4.3 进行计算，这里不再举例。

4. 斯皮尔曼(Spearman)等级相关系数 r_s

当两个变量测量得到的数据均以等级形式呈现，或者测量得到的数据是非正态分布时，两个变量间的相关问题使用斯皮尔曼等级相关方法进行计算，计算公式如下：

$$r_s=1-\frac{6\sum_{i=1}^{n}D_i^2}{n(n^2-1)}。\tag{4.4}$$

其中，D_i 表示两个变量每对数据的等级之差；n 表示样本容量。

【例 4－3】 某 10 名小学生参加数学竞赛、单元测验等的两次测验成绩如表 4－5，试求两次测验间的相关系数。

表 4－5　10 名小学生两次数学单元测验成绩表

学号	竞赛	单元	竞赛等级	单元等级	等级差 D	D^2
1	63	98	9	3	6	36
2	92	97	2.5	4.5	−2	4
3	88	96	5	6	−1	1
4	89	95	4	7.5	−3.5	12.25
5	74	95	7.5	7.5	0	0
6	96	99	1	1.5	−0.5	0.25
7	49	97	10	4.5	5.5	30.25
8	74	93	7.5	9	−1.5	2.25
9	92	99	2.5	1.5	1	1

（续表）

学号	竞赛	单元	竞赛等级	单元等级	等级差 D	D^2
10	81	87	6	10	−4	16
合　计			55	55	0	103

分析：由于不能确定两次测验的成绩是否服从正态分布，因此采用斯皮尔曼(Spearman)等级相关方法进行计算。

解法 1：将表 4－5 中的两次测验成绩由高到低进行排序，遇到两个分数相同时，用它们所占等级位置的平均数作为它们的等级。

$$r_s = 1 - \frac{6 \times 103}{10 \times (10^2 - 1)} \approx 0.376。$$

即这 10 名学生参加数学竞赛的成绩与单元测验成绩之间的等级相关系数是 0.376<0.5，说明竞赛成绩与单元测验成绩的一致性程度不高。

解法 2：本题如果用 SPSS 软件计算，共分 4 步，颇为方便。

1. 将数据导入 SPSS 数据编辑器的工作表中。

2. 如图 4－1，执行【分析】/【相关】/【双变量】程序，出现“双变量相关”对话框。

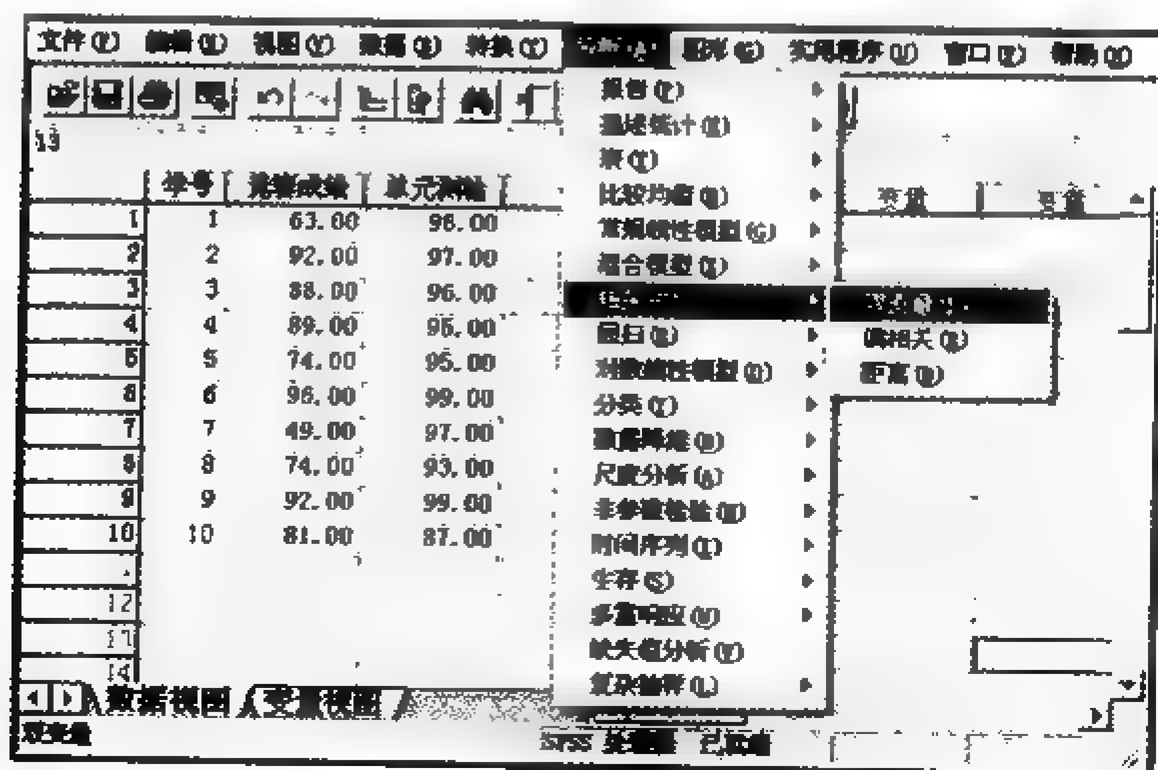


图 4－1

3. 如图4 2,将左边方框中的题目竞赛成绩与单元测验选入右边的“变量”下的空框中。在“相关系数”选项下,选中“Spearman”;在“显著性检验”选项下,选中“双侧检验”;选中“标记显著性相关”。

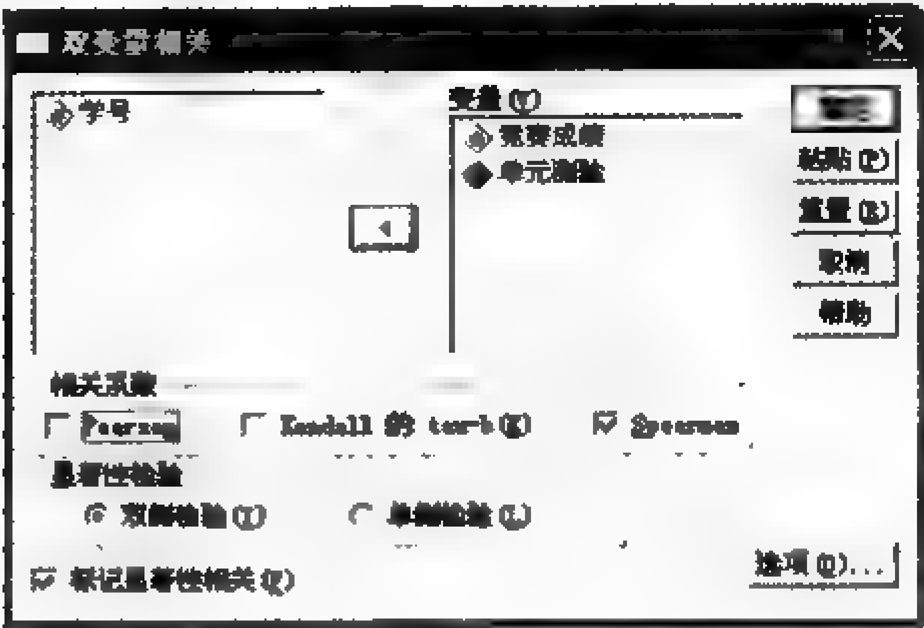


图 4-2

4. 单击“确定”按钮,执行程序计算。输出结果如表4-6。

表 4-6 竞赛成绩与单元测验成绩的相关性

			竞赛成绩	单元测验
Spearman 的 rho	竞赛成绩	相关系数 显著性(双侧) N	1.000 10	.366 .298 10
	单元测验	相关系数 显著性(双侧) N	.366 .298 10	1.000 10

表4-6显示,样本总数是10,竞赛成绩与单元测验的Spearman相关系数值为0.366,双尾检验的p值为0.298,可以接受零假设“竞赛成绩与单元测验无关”。

需要注意的是,两种解法得出的Spearman相关系数值略有出入,这跟计算时的精确度处理有关,但差异值很小(0.376与

0.366)，并不影响结论的判断。

三、相关系数的应用

考试研究中，相关分析具有较大的实用价值，它既可以帮助研究者对考试的质量与效果做出正确决策，也可以帮助学校与教师正确认识学生的掌握情况以改进教学与管理。进行试卷质量分析时，相关分析主要应用于两个方面。

1. 用于分析试卷的整体质量

由于试卷的测验信度、测验效度的计算都需要用到积差相关系数，除了计算有关的相关系数外，还需要结合测量学理论、试卷的答题情况进行深入分析与判断，具体详见第二章相关章节。

2. 用于分析试题质量

试题的难度与区分度的计算需要用到点二列相关、二列相关、积差相关等方法，需要根据每种计算方法的适用条件、计算结果精确度的要求以及研究的目的等多种因素加以确定。作相关分析时，一般而言，能够用“积差法”计算相关系数的，就不要用“等级法”计算相关系数，以免失掉原始数据中的一些有价值的信息。

3. 注意区分相关显著性水平与相关密切程度

相关显著性水平通常分为两种：有显著性（$p \leqslant 0.05$）或有高度显著性（$p \leqslant 0.01$），其意义是由样本信息推断两个变量所属总体也呈相关的可能性有95%或99%，即由于抽样误差的原因，还有5%或1%的推断不准的可能性（两个变量实际上并不相关）。也就是说，相关显著性水平研究的是两个变量之间有无相关的问题。相关显著性水平的检验方法参见一般的教育统计学书籍。

相关密切程度是由相关系数$|r|$来度量的，一般而言，$|r|$越大，表明两个变量关系越密切，相关程度越高；$|r|$越小，两个变量相关程度越低。

第二节　回归分析

在教育研究实践中，很多研究对象之间存在着相关关系，如，代数学习与几何学习的关联，性别与数学学习的关联等。相关分析反映了两个变量之间是否有关、互相关联的密切程度，但是它并没有揭示出两个变量之间是如何相互依存、相互影响与发展变化的。由于变量间的这种关联具有不确定性，不能用数学中的函数关系来表示，因此需要借助统计学中有关的理论与方法做进一步探讨。在统计学中，研究变量间相互依存变化的规律，以便依据已知变量值估计预测未知变量值的计算理论与方法称为回归分析。

一、回归分析与相关分析

回归分析和相关分析是研究变量间非确定性关系的两个重要工具，在应用中，两种分析方法互相渗透、相互结合。

1. 回归与相关的联系

回归与相关都是研究变量间的关联问题。利用回归分析，可以建立两个变量之间的函数关系，然后借助函数关系，由一个变量值来估计、预测另一个变量值，估计与预测的前提是两个变量间存在相关关系。如果两个变量之间相关为0，即两个变量无关联，那么即使求出回归方程，也不能由一个变量值估计预测另一个变量值，此时的回归无意义。在存在相关的情况下，两个变量的相关程度越高，由一个变量值估计预测另一个变量值时，所得出的结论就越可靠。

2. 回归与相关的区别

回归与相关的区别主要在于两者的研究角度不同，具体体现在以下三个方面。

（1）变量关系的单一性

相关分析研究的是变量间是否存在相互关系、互相关联的程度如何，两个变量处于平等的地位，即使变量间存在相关关系，也可能存在多种关联方式，不一定就是因果关系。

回归分析是通过建立回归方程，依据自变量的已知值去预测因变量的未知值，因此，两个变量的地位并不平等，其中因变量处于被解释的地位。

（2）变量取值的随机性

相关分析中，两个变量都是随机变量，每个变量的取值都具有随机性；但回归分析中，因变量是随机变量，但自变量可以是随机的，也可以是研究者指定选择的。例如，研究中学生的记忆力与数学学习效果的关系，想求出反映记忆力与数学学习效果关系的回归方程，这时研究者可以选取一定数量具有某种记忆力水平的中学生被试，这时，自变量（中学生被试的记忆力水平）取值是非随机的变量，而因变量（对应被试的数学测验成绩）是随机的。

（3）变量研究的对称性

在相关分析中，变量 x 与变量 y 的相关具有双向对称性，即 x 与 y 的相关和 y 与 x 的相关是一样的。但在回归分析中，把 x 作为自变量、y 作为因变量和把 y 作为自变量、x 作为因变量，得到的回归方程是不同的，即回归分析具有不对称性。

二、回归分析的主要步骤与基本类型

1. 主要步骤

利用回归分析方法研究变量之间的关系，主要分三步进行：构建回归方程，检验和评价所建立的回归方程的有效性，利用所构建的回归方程进行估计、预测与控制。

（1）构建回归方程

构建回归方程是回归分析的核心部分。根据测验得到的数据，先制作散点图，观察变量间相关方式与相关程度；然后选择适当的函数关系式作为回归方程的模型，确定自变量的个数；再遵循最佳

拟合原则，估计回归方程中的所有参数。得出所有参数的估计值后，即构建出变量间的回归方程。

(2) 检验和评价所构建的回归方程的有效性

理论上，不论变量间是否存在相关关系，也不论自变量个数、样本大小如何，按照第1步的思路均可以构建出变量间的回归方程。现在需要检验构建的回归方程是否有用？如果有用，那么如何评价使用价值的高低？如果有效性低，是否可以修正回归方程？可以从哪些方面修正回归方程？对这些问题的研究，形成回归分析的第2部分。

(3) 利用所构建的回归方程进行估计、预测与控制

经过检验确定为有效的回归方程，就可以用来对因变量进行估计、预测或控制。例如，估计或预测因变量的取值范围，估计或预测因变量的关键取值，利用回归方程揭示变量间的关系，通过控制或调整自变量的取值而达到控制因变量变化趋势的目的，等等。

2. 基本类型

回归分析的基本类型是由变量个数、变量类型、变量之间的相关关系和选择的函数类型决定的。

(1) 一元回归与多元回归

按照回归方程中涉及自变量个数的多少，可以将回归分析分成一元回归与多元回归两种。只研究两个变量之间的回归关系的，称为一元回归分析；研究2个或2个以上自变量与因变量之间的回归关系的，则称为多元回归分析。例如，研究小学毕业数学成绩对初中毕业数学成绩的预测功能时，属于一元回归分析；研究试题考查的能力层次、内容深度、考查方式对试题难度的影响时，需要使用三元回归分析。

(2) 线性回归与曲线回归

按照构建回归方程选用的函数模型的不同，可以将回归分析分成线性回归和曲线回归两种类型。在实际应用中，变量中的一部分

曲线相关关系可以通过一定的数学变换转化为线性相关关系，从而利用线性回归分析。

回归分析中所应用的模型多种多样，但以线性回归模型的应用最为广泛。在教育测验中，最常见的是一元线性回归和二元线性回归，下面分别介绍一元线性回归和二元线性回归的原理与具体操作方法。

三、一元线性回归

一元线性回归分析是在排除其他影响因素或假定其他影响因素确定的前提下，研究某一个因素（自变量）对另一个因素（因变量）的影响过程。这种分析突出抓主要因素的特点，带有理想化的成分。

一元线性回归分析只涉及两个变量 x 与 y，习惯上，称 y 为因变量，x 为自变量。由于假设变量 x 与 y 之间为线性关系，因此，用一次函数模型来构建一元线性回归方程。

1. 一元线性回归方程及其求法

从一个具体例子谈起。

【例 4－4】 表 4－7 是某学校 20 名学生同一学年两个学期期末数学测验成绩，其中 x 表示上学期期末数学测验成绩，y 表示下学期期末数学测验成绩。试确定两次测验成绩的线性关系，为下一届学生的成绩预测做准备。

表 4－7　20 名学生两个学期期末数学测验成绩

序号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
x	80	60	82	72	73	80	86	90	85	78	95	72	80	75	60	62	50	65	70	90
y	76	66	80	75	69	85	90	84	93	76	93	76	78	81	55	68	57	62	77	85

利用 EXCEL 软件的作图功能，得到 x 与 y 这两组数据的散点图，如图 4－3。这 20 个点大致分布在一条直线附近，因此可以用一次函数 $y=a+bx$ 来近似反映变量 x 与 y 之间的关系。

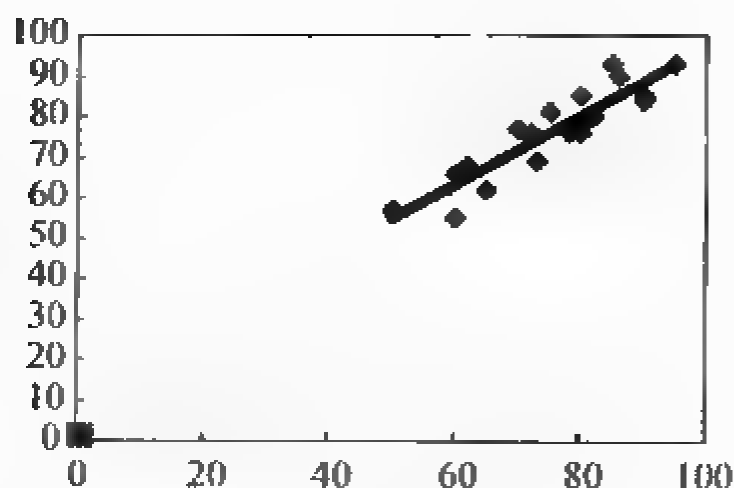


图 4-3

(1) 一元线性回归模型

一元线性回归方程的回归模型为

$$y = a + bx。\quad (4.5)$$

其中，y 为因变量，x 为自变量，a、b 为待定系数。

从图 4-3 可以看到，可以作出不同的直线，使得数据对应的点都在这些直线附近。也就是说 a、b 的取值可以有不同的方法。这就涉及选择最佳方案问题。

(2) 回归方程构建原理

依据图 4-3 来分析。

如果点 (x_1, y_1) 在直线 $y = a + bx$ 上，那么 $y_1 = a + bx_1$，即 $y_1 - a - bx_1 = 0$。即函数 $y = a + bx$ 准确地刻画出 x_1、y_1 之间的关系。

如果点 (x_1, y_1) 不在直线 $y = a + bx$ 上，那么 $y_1 \neq a + bx_1$，令误差 $\varepsilon_1 = y_1 - a - bx_1$，显然误差 ε_1 越小越好。为了消除正负号的影响，人们用 ε_1^2 来做进一步研究。

综合所有的这 20 个数据点，当所有误差的平方和 $\sum \varepsilon_i^2 = \sum (y_i - a - bx_i)^2$ 最小时，可以求出参数 a、b 的值，此时得到的直线是拟合最好的。

这就是建立回归方程的基本原理。这种方法称为最小二乘方法，它是确定回归直线的最有力的工具。

(3) 回归系数计算公式

根据最小二乘方法，对 a、b 求偏导，可以得出关于 a、b 的二元一次方程组，进而得出关于 a、b 的计算公式如下：

$$b=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2},\tag{4.6}$$

$$a=\bar{y}-b\bar{x}。\tag{4.7}$$

其中，x_i、y_i 是成对的测验分数，$\bar{x}$、$\bar{y}$ 是相应的测验成绩平均数。

我们把这种能反映变量 x、y 之间的最佳拟合关系的直线方程称为回归直线方程，通常记作 $\hat{y}=a+bx$，其中用 $\hat{y}$ 代替 y 是表明 $\hat{y}=a+bx$ 仅仅是对 x、y 之间确定关系的一种估计。人们也称 b 为回归系数，它表示回归直线的斜率，a 是回归直线在纵轴上的截距。

a、b 的计算可以用公式解决，但是用 EXCEL 软件与 SPSS 软件都可以很快完成。

例 4-4 解答：下面简要介绍用 EXCEL 软件求例 4-4 中两次测验成绩之间的线性回归方程。

① 将表 4-7 中的成绩输入 EXCEL 工作表，如图 4-4。

	A	B	C	D	E	F
1	学号	x	y			
2	1	80	76			
3	2	60	66			
4	3	82	80			
5	4	72	75			
6	5	73	69			
7	6	80	85			
8	7	86	90			
9	8	90	84			
10	9	85	93			
11	10	78	76			
12	11	95	93			
13	12	72	76			

例4-4

图 4-4

② 横向并列选取两个单元格 D2、E2，用以输入计算所得的待定参数 a、b 的值。

③ 执行【插入】/【函数】程序，出现“插入函数”对话框，如图 4-5，在“选择类别”后方框中选中“统计”，在“选择函数”栏目下选中“LINEST”（即计算线性回归方程的参数）。单击“确定”按钮，出现“函数参数”对话框。



图 4-5

④ 如图 4-6，在 Known_y's 后的方框中输入“C2:C21”（即因变量的取值），在 Known_x's 后的方框中输入“B2:B21”（即自变量的



图 4-6

取值)。在按住“Shift＋Ctrl”键的同时，单击“确定”按钮，则在 D2、E2 单元格中分别显示出参数 a、b 的值，如图 4－7。

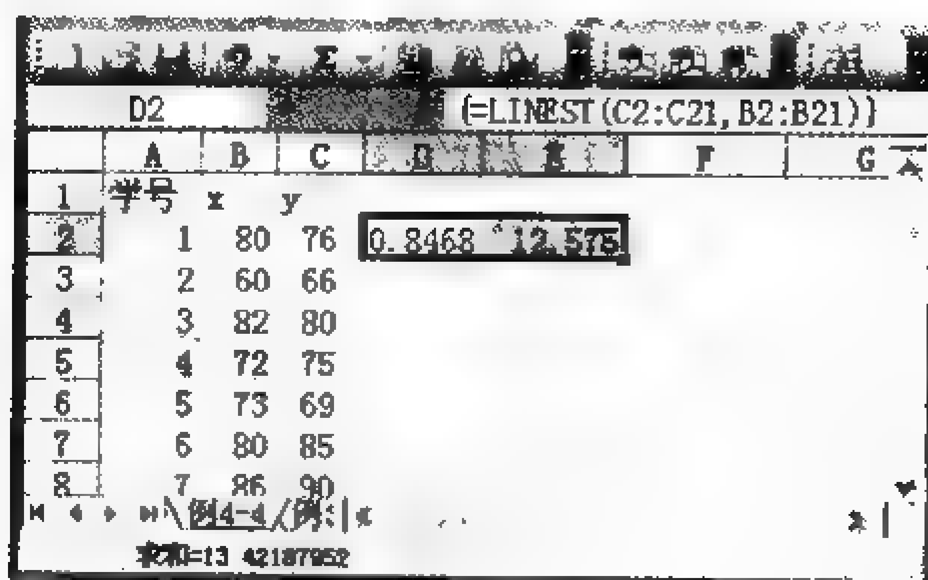


图 4－7

所以，两次测验成绩之间的线性回归方程为 $\hat{y} = 0.8468x + 12.575$。

另外，也可以利用 EXCEL 软件中的作图功能完成回归直线的作图。

① 先利用图表向导完成散点图。

② 双击选中散点图，在散点上鼠标右击，出现图表选项，如图 4－8。选择“添加趋势线”，即可出现图 4－3 所示的回归直线。

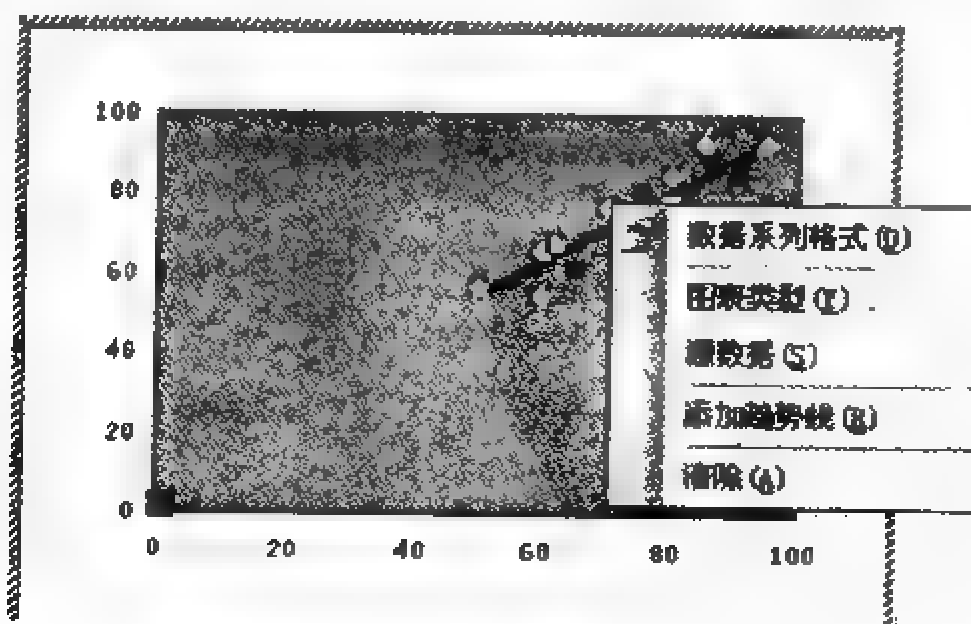


图 4－8

2. 一元线性回归方程的有效性检验

根据原始数据求出回归方程后，由于这一回归直线是根据有限的样本信息获得的，它是否反映出两个变量间的真实联系规律，需要进行检验。另外，还需要检验回归方程对因变量的预测效果。

(1) 方差分析，判定回归方程是否有效

在回归分析中，因变量 y 与平均数 $\bar{y}$ 的差异的偏差平方和 $SS_t = \sum(y_i - \bar{y})^2$ 可以分解成回归平方和与误差平方和两部分，即

$$SS_t = SS_r + SS_e。 \tag{4.8}$$

其中 $SS_r = \sum(\hat{y}_i - \bar{y})^2$ 是回归平方和，它完全由自变量 x 所决定，反映的是 x 的重要程度；$SS_e = \sum(y_i - \hat{y}_i)^2$ 是残差平方和，它反映的是随机因素对因变量造成的影响。

回归方程有效性检验的零假设是"所求回归方程无效"，也即"回归系数值为 0"，假设的实质是由自变量决定的回归方差并不显著大于残差方差。一元线性回归方差分析的总误差平方和的自由度为 $n-1$，回归平方和的自由度为 1，残差平方和的自由度为 $n-2$，统计量 $F = \dfrac{SS_r/1}{SS_e/(n-2)}$ 服从自由度为$(1, n-2)$的 F 分布。在显著性水平 α 确定的条件下，根据回归自由度 1 与残差自由度 $n-2$，查 F 分布表，可以得到检验临界值 $F\alpha$，如果 $F > F\alpha$，则拒绝零假设，说明有 $1-\alpha$ 的把握确定所求方程有效，可以实际使用；反之，则接受零假设，说明所求方程无效。

例 4-4 解法(续 1)：下面，对例 4-4 中求出的两次测验成绩之间的线性回归方程 $\hat{y} = 0.8468x + 12.575$ 是否有效进行检验。也借助 EXCEL 软件完成。

① 执行【插入】/【函数】程序，出现"插入函数"对话框；在"选择类别"后方框中选中"统计"，在"选择函数"栏目下选中"LINEST"，单击"确定"按钮，出现"函数参数"对话框。在 Known_y's 后的方框中输入"C2：C21"，在 Known_x's 后的方框中输入"B2：B21"，在

Const 后的方框中输入"true",在 Stats 后的方框中输入"true"。如图 4－9。

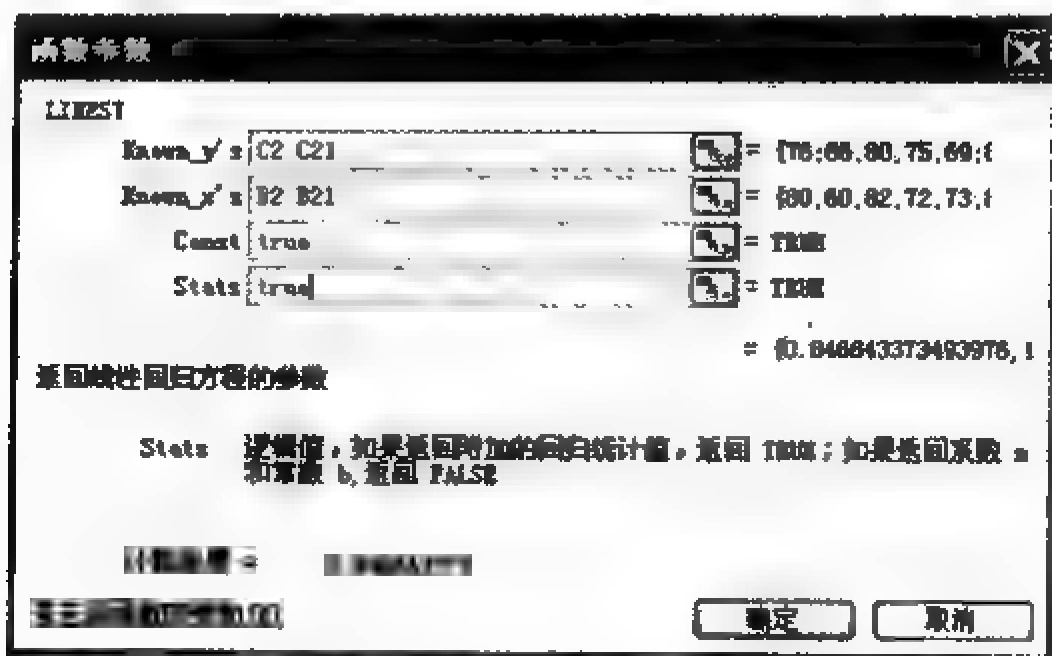
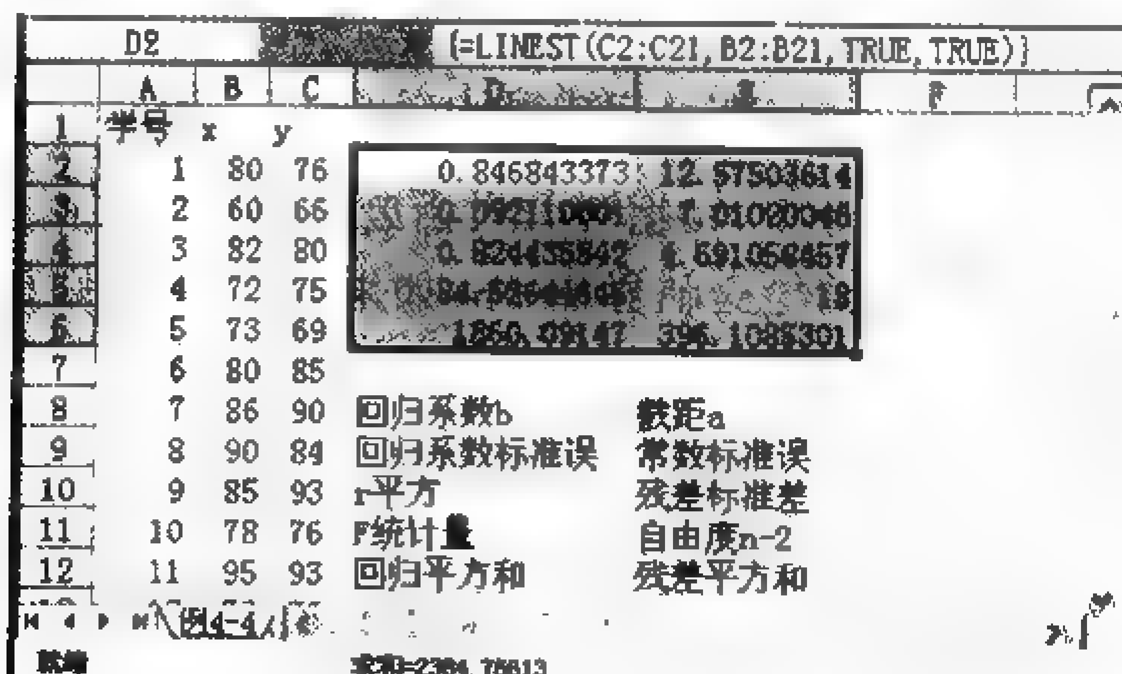


图 4－9

② 在按住"Shift＋Ctrl"键的同时,单击"确定"按钮,则在单元格 D2:E6 区域中分别显示出 10 个统计量的值,如图 4－10,每个统计量的含义在单元格 D8:E12 区域中相对应的位置给予了解释。



D2　{=LINEST(C2:C21,B2:B21,TRUE,TRUE)}

	A	B	C	D	E
1	学号	x	y		
2	1	80	76	0.846843373	12.5750361?
3	2	60	66	[illegible]	7.01020046
4	3	82	80	0.824435842	4.691058457
5	4	72	75	[illegible]	18
6	5	73	69	1860.09147	396.1085301
7	6	80	85		
8	7	86	90	回归系数b	截距a
9	8	90	89	回归系数标准误	常数标准误
10	9	85	93	r平方	残差标准差
11	10	78	76	F统计量	自由度n-2
12	11	95	93	回归平方和	残差平方和

图 4－10

③ 如图 4－11,取显著性水平 α　0.01,利用函数"FINV"计算临界值 $F_{0.01}(1, 18)$。在单元格 F2 中输入"＝FINV(0.01, 1,

18)”,按 Enter 键,返回值 8.285419545 就显示在单元格 F2 中,即 $F_{0.01}(1,18)=8.285419545$。

$\because F=84.52644645>F_{0.01}(1,18)$,

$\therefore$ 有 99%的把握说两次测验成绩之间存在线性关系。

F2 =FINV(0.01,1,18)

	A	B	C	D	E	F
1	学号	x	y			显著性水平0.01
2	1	80	76	0.846843373	12.57503614	8.285419545
3	2	60	66	0.092110001	7.01020046	
4	3	82	80	0.824435542	4.691058457	
5	4	72	75	84.52644645	18	
6	5	73	69	1860.09147	396.1085301	
7	6	80	85			
8	7	86	90	回归系数b	截距a	
9	8	90	84	回归系数标准误	常数标准误	
10	9	85	93	x平方	残差标准差	
11	10	78	76	F统计量	自由度n-2	
12	11	95	93	回归平方和	残差平方和	

例4-4

图 4-11

(2) 决定系数,衡量回归方程有效性的高低

回归分析中衡量回归方程有效性高低的指标称为决定系数,记作 R^2,其计算公式为:

$$R^2=\frac{SS_r}{SS_t}。\tag{4.9}$$

由公式 4.9 知,R^2 是回归平方和在总偏差平方和中所占的比例。经过推理,可以进一步发现 $R^2=r_{xy}^2$,即在一元线性回归中,决定系数 R^2 是因变量 y 与自变量 x 积差相关系数的平方。因此,可以说回归分析是相关分析的继续与发展,回归分析对回归方程有效性的检验实质上是对变量相关显著性的检验。

例 4-4 解法(续 2):下面,对例 4-4 中求出的两次测验成绩之间的线性回归方程 $\hat{y}=0.8468x+12.575$ 有效性的高低进行检验。也借助 EXCEL 软件完成。

在图 4-11 中,已经求出决定系数 $R^2\approx0.8244$,说明因变量 y 离差中由自变量 x 所决定的部分占 82.44%,因此,回归方程有效程

度高。

需要特别注意的是，经过方差分析检验回归方程无效后，再求取决定系数是无意义的。

3. 一元线性回归分析的估计与预测

求回归方程的目的主要是利用回归方程对因变量进行估计与预测。在应用一元线性回归方程进行估计与预测时，主要有以下两种情况。

(1) 用样本回归方程估计因变量回归值 $\hat{y}$ 的变化范围

对于自变量 x 的一个确定值，因变量 y_i 仍然是随机的。在实际问题中，我们希望能够估计在 x 确定的前提下，因变量回归值 $\hat{y}$ 的变化范围。表 4－8 给出了确定 $\hat{y}$ 变化范围的计算方法与计算公式。

表 4－8　根据 x_0 确定 $\hat{y}_0$ 变化范围的计算方法与计算公式

显著性水平	大样本(正态分布，$n \geq 30$)		小样本(t 分布，$n < 30$)	
	下限	上限	下限	上限
0.1	$\hat{y}_0 - 1.64s_{yx}$	$\hat{y}_0 + 1.64s_{yx}$	$\hat{y}_0 - t_{0.1(df)}s_{yx}$	$\hat{y}_0 + t_{0.1(df)}s_{yx}$
0.05	$\hat{y}_0 - 1.96s_{yx}$	$\hat{y}_0 + 1.96s_{yx}$	$\hat{y}_0 - t_{0.05(df)}s_{yx}$	$\hat{y}_0 + t_{0.05(df)}s_{yx}$
0.01	$\hat{y}_0 - 2.58s_{yx}$	$\hat{y}_0 + 2.58s_{yx}$	$\hat{y}_0 - t_{0.01(df)}s_{yx}$	$\hat{y}_0 + t_{0.01(df)}s_{yx}$

其中 $s_{yx} = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2}$ 是回归直线的残差标准差。

【例 4－5】 在例 4－4 中，已经建立某校下学期期末数学测验成绩对上学期期末数学测验成绩的一元线性回归方程 $\hat{y} = 0.8468x + 12.575$，残差的标准差约为 4.69，经检验回归方程有效，且有效性高。假设学生甲上学期期末数学测验成绩 $x = 80$ 分，求该生下学期期末数学测验预测成绩 y 的 95%正常范围。

解：由于样本数 $n = 20$，是小样本情况，因此属于 t 分布。

当 $x = 80$ 时，$\hat{y} = 0.8468 \times 80 + 12.575 \approx 80.32$。

$df = n - 2 = 18$，$\alpha = 0.05$，查 t 值表得 $t_{0.05(18)} = 2.101$，则

下限为 $80.32 - 2.101 \times 4.69 = 70.47$，上限为 $80.32 + 2.101 \times$

4.69 = 90.17。

所以，该生下学期期末数学测验预测成绩 y 的 95% 正常范围为 [70.47, 90.17]。

（2）根据样本因变量回归值 $\hat{y}$ 预测因变量真值 y 的置信区间

利用回归方程由自变量 x 的值在一定概率意义下估计出因变量 y 的取值范围，这个估计范围只考虑了 y 值在回归直线上下波动，并没有考虑回归直线本身的波动。由于回归方程建立在样本数据上，不同的样本构建的回归方程不一定相同，因此用现有的回归方程计算出的回归值，不一定就是因变量的真实值。要想求出因变量的真实值，就需要用到样本方程的误差标准差，再根据正态分布，对因变量真实值的置信区间做出概率估计。有兴趣的读者可以查阅有关的统计学书籍①。

四、多元线性回归

一元线性回归只研究一个自变量对因变量的影响过程，是回归分析中最简单的情况。在实际问题中，影响因变量的因素往往多于1个，例如学生学业成绩既受学生的智力水平影响，还受个人学习方法、教师教学方式等因素的影响，这就需要研究因变量与多个自变量的关系，研究工具就是多元回归分析。

多元线性回归分析的原理与一元线性回归分析的原理相同，但在具体计算上却复杂得多，人们往往借助计算机来完成。下面主要介绍多元线性回归方程的建立与检验。

1. 多元线性回归方程及其求法

（1）多元线性回归模型与参数意义

多元线性回归模型为

$$y = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k。 \qquad (4.10)$$

---

① 例如，凌云著，考试统计学，武汉：华中师范大学出版社，2006 年 12 月第 2 次印刷。

其中，a 是常数项，$b_1, b_2, \cdots, b_k$ 分别是自变量 $x_1, x_2, \cdots, x_k$ 的回归系数，也简称为偏回归系数。a 表示的是在所有自变量都保持不变的情况下，因变量 y 的平均变化率。b_i 表示的是在其他自变量都保持不变的情况下，自变量 x_i 每变化一个单位，因变量 y 的平均变化率。

(2) 多元线性回归方程构建原理

建立多元线性回归方程同样是利用最小二乘方法，在使得回归估计值 $\hat{y}_i$ 与实测值 y_i 的误差平方和最小，即 $\sum \varepsilon_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - a - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_k x_{ik})^2$ 取得最小值时，利用求偏导数的方法分别求出各个参数 $a, b_1, b_2, \cdots, b_k$ 的值。

由于建立多元线性回归方程所使用的数据仍然只能是样本数据，因此，所建立的方程还是样本回归方程，通常记为

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k。 \quad (4.11)$$

2. 多元线性回归方程的有效性检验

(1) 方差分析

多元线性回归方程建立后，同样必须经过统计检验才能判断它是否具有实用价值，检验的方法还是方差分析法。回归方程有效性检验的零假设是“所求回归方程无效”，多元线性回归方差分析的总误差平方和的自由度为 $n-1$，回归平方和的自由度为自变量的个数 k，残差平方和的自由度为 $n-k-1$，统计量 $F = \dfrac{SS_r/k}{SS_e/(n-k-1)}$ 服从自由度为 $(k, n-k-1)$ 的 F 分布。在显著性水平 α 确定的条件下，根据回归自由度 k 与残差自由度 $n-k-1$，查 F 分布表，可以得到检验临界值 $F\alpha$，如果 $F > F\alpha$，则拒绝零假设，说明有 $1-\alpha$ 的把握确定所求方程有效，可以实际使用；反之，则接受零假设，说明所求方程无效。

(2) 决定系数与复相关系数

多元线性回归分析中衡量回归方程有效性高低的指标仍然是

决定系数 R^2，其计算公式仍为 $R^2 = SS_r/SS_t$。另外，决定系数 R^2 的算术根 R 表示因变量 y 与自变量 $x_1, x_2, \cdots, x_k$ 之间的相关程度，因此，R 又称为 y 与 $x_1, x_2, \cdots, x_k$ 的复相关系数，它是衡量样本观测值与回归估计值之间线性程度的指标。

（3）偏回归系数显著性检验

多元线性回归中，方程的显著性检验是检验多个自变量共同对因变量的影响是否显著，即使影响显著，也不能说明每个自变量对因变量都有显著影响。因为影响显著，只是说明方程中有 1 个或多个偏回归系数不为 0，并不是说每一个偏回归系数都不等于 0。因此，还必须对每一个自变量的回归系数进行统计检验，以确定每个自变量对因变量是否有影响。

对每个偏回归系数的显著性检验也都通过计算机来完成。

对自变量进行检验和筛选后，应剔除那些对因变量没有影响或影响很小，经检验未达到显著性水平、不足以入选的自变量，以达到简化变量间关系结构和所求回归方程的目的。因此，最终求得的多元线性回归方程入选的自变量个数可能少于最初选择的数目。

利用 SPSS 软件进行多元线性回归分析的实例参见例 4－12。

第三节　方差分析

在实际研究中，人们经常遇到需要对不同学校的统考成绩进行比较，看不同学校之间的考试结果是否有差异；或者，对同一学校同一个年级不同班级的测验成绩进行比较，看看不同班级之间的考试结果是否有差异。这类研究的共同特点是对同一个变量（某个测验成绩）的多个总体（不同学校、不同班级等）的平均值进行比较，比较采用的方法就是方差分析法（analysis of variance，简称 ANOVA）。在第二节进行线性回归方程的有效性检验时，我们已经使用了方差分析法，下面具体介绍方差分析的基本原理与简单应用。

一、方差分析的基本原理

方差分析又称为变异数分析，它是英国统计学家 Fisher 首先提出的一种统计方法，因此有时也称为 F 检验。方差分析的基本思想是把得到的所有观测数据分成几个组，分析数据中不同来源的变异对总体变异的贡献大小，从而确定自变量对因变量的影响是否显著。

1. 方差分析的逻辑基础

从一个具体例子谈起。

【例 4－6】 某校初三年级六个班共 299 名学生。初三上学期期末测验数学成绩输入 EXCEL 工作表后如图 4－12(满分 150 分)，请对初三年级该次测验成绩进行分析与比较。

	A	B	C	D	E	F	G	H	I
1	班级	学号	数学成绩		班级	平均分			
2	1	1	148		1	136.76			
3	1	2	146		2	136.14			
4	1	3	118		3	79.1			
5	1	4	135		4	78.52			
6	1	5	131		5	83.94			
7	1	6	133		6	86.8			
8	1	7	145		总计	100.09			
9	1	8	142						
299	6	298	128						
300	6	299	53						

图 4－12

由图 4－12 知，初三年级六个班数学平均成绩之间有差异，而每个班内各个学生成绩之间也有差异。全年级的学生数学成绩之间存在很大差异，这种差异可以大致分成两种来源：班与班之间的差异，班内学生成绩间的差异，通常把前者称为样本组间差异，把后者称为样本组内差异。如果组间差异占较大比例，则认为班与班之间的教学效果差异明显；如果组内差异占较大比例，则认为全年级的成绩差异主要由随机误差造成，班与班之间的教学效果差异不明显。

在方差分析中，以收集的所有数据与总平均数的偏差平方和作为变异的统计量，表示为

$$SS_{总} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_t)^2。 \quad (4.12)$$

其中，$SS_{总}$ 表示总变异（总平方和），x_{ij} 表示每个数据值，$\overline{x}_t$ 表示总平均数，k 表示数据组数，n_i 表示第 i 组的数据个数。

总变异 $SS_{总}$ 可以分解成组间变异（组间平方和）与组内变异（组内平方和）两部分，即

$$SS_{总} = SS_{组间} + SS_{组内}。 \quad (4.13)$$

其中组间变异 $SS_{组间} = \sum_{i=1}^{k} n_i(\overline{x}_i - \overline{x}_t)^2$ 是各个组的样本平均数与总平均数偏差的平方和，反映的是分组因素的重要程度；

组内差异 $SS_{组内} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)^2$ 是每个组内各个数据与本组平均数偏差的平方和，它反映的是随机因素对因变量造成的影响。

在方差分析中，不能直接比较 $SS_{组间}$ 与 $SS_{组内}$ 的大小，因为它们与数据的个数、分组个数等有关。为了消除个数的影响，往往求其方差，又称为均方，即偏差平方和除以各自的自由度。其中，组间均方 $MS_{组间} = \frac{SS_{组间}}{df_{组间}}$，组内均方 $MS_{组内} = \frac{SS_{组内}}{df_{组内}}$。研究者关心的是组间均方是否显著地大于组内均方，因此，在求 F 值时，把组间均方放在分子上，即 $F = \frac{MS_{组间}}{MS_{组内}}$，采取单侧检验。如果 $F \leqslant 1$，说明数据的总变异中组间变异所占比例小于或等于组内变异，认为组间差异不显著；如果 $F > 1$，说明数据的总变异中组间变异所占比例大于组内变异，如果满足 $F > F_a(df_{组间}, df_{组内})$，则认为组间差异显著。

2. ***方差分析的基本条件***

应用方差分析时，数据必须满足三个基本假定。

（1）变异的可分解性

变异的可分解性是进行方差分析依据的基本原理，即要求数据的总变异能够被分解成几个部分，每个部分的变异来源意义明确、相

互独立。例如,例 4-6 中的测验成绩总变异可以分解为组间变异和组内变异,回归分析中总变异分解为回归平方和与残差平方和等。

(2) 总体服从正态分布

进行方差分析时,要求各个样本来自的总体呈正态分布。在教育测验中,测验成绩大多数满足正态分布要求,一般不需要进行正态性检验。如果已经认定样本来自的总体属于非正态分布,就应该将数据进行正态性转换,或采用非参数检验法。

(3) 方差齐性

进行方差分析时,要求各个样本来自的总体方差相等,即要求方差齐性,否则不能进行方差分析。一般地,在进行方差分析之前,要用哈特莱(Hartley)最大 F 值法对样本方差进行齐性检验,以便确定能否进行方差分析。

二、单因素方差分析

1. 实施单因素方差分析的前提条件

例 4-6 中分析全年级数学成绩时,只考虑班级这一个因素,我们称它为因素 A,而全年级分成的六个班称为因素 A 的 6 个水平,即 6 个不同取值。类似地,单因素方差分析的前提条件可以表示成表 4-9。

表 4-9 单因素方差分析的前提条件

		样本	容量	平均数
因素 A	水平 1	$x_{11}, x_{12}, \cdots, x_{1,n1}$	$n1$	$\overline{x}_1$
	水平 2	$x_{21}, x_{22}, \cdots, x_{2,n2}$	$n2$	$\overline{x}_2$
	⋮	⋮	⋮	⋮
	水平 k	$x_{k1}, x_{k2}, \cdots, x_{k,nk}$	nk	$\overline{x}_k$

2. 单因素方差分析的一般步骤

(1) 建立零假设

首先,随机抽取 k 个样本(即因素 A 的 k 个水平),它们均来自具有相同方差的正态总体。然后,提出零假设"k 个总体的平均数都

相等”，那么备择假设即“至少有 2 个总体的平均数不相等”。

（2）计算统计量并建立方差分析表

利用公式 4.12 与 4.13 等，计算 $SS_{总}$、$SS_{组间}$、$SS_{组内}$、$MS_{组间}$、$MS_{组内}$、F 值，并填写在表 4－10 中。

表 4－10　方差分析表

变异来源	平方和	自由度	均方(方差)	F 值	F 值右尾概率 P（显著性）
组　间	$SS_{组间}$	$k-1$	$MS_{组间}$	$MS_{组间}/MS_{组内}$	
组　内	$SS_{组内}$	$n-k$	$MS_{组内}$		
总　体	$SS_{总}$	$n-1$			

（3）进行统计推断

根据表 4－10，如果 F 值达到 0.05 显著水平的临界值，则说明各平均数间差异显著；若 F 值达到 0.01 显著水平的临界值，则说明各平均数间差异非常显著。

例 4－6 的解答：下面运用 SPSS 软件中的单因素方差分析，分析初三年级某次测验六个班级之间平均数是否有差异，差异是否显著，共分四步。

第 1 步，如图 4－13，将图 4－12 中的数据导入 SPSS 数据编辑

	班级	考生号	sum	变量
1	1	1	148	
2	1	2	146	
3	1	3	118	
4	1	4	136	
5	1	5	131	
6	1	6	133	
7	1	7	145	
8	1	8	142	
9	1	9	134	
10	1	10	136	
11	1	11	141	
12	1	12	145	

数据视图　变量视图　　SPSS 处理器

图 4－13

器的工作表中。

第 2 步，如图 4－14，执行【分析】/【比较均值】/【单因素ANOVA】程序，出现“单因素 ANOVA”对话框。

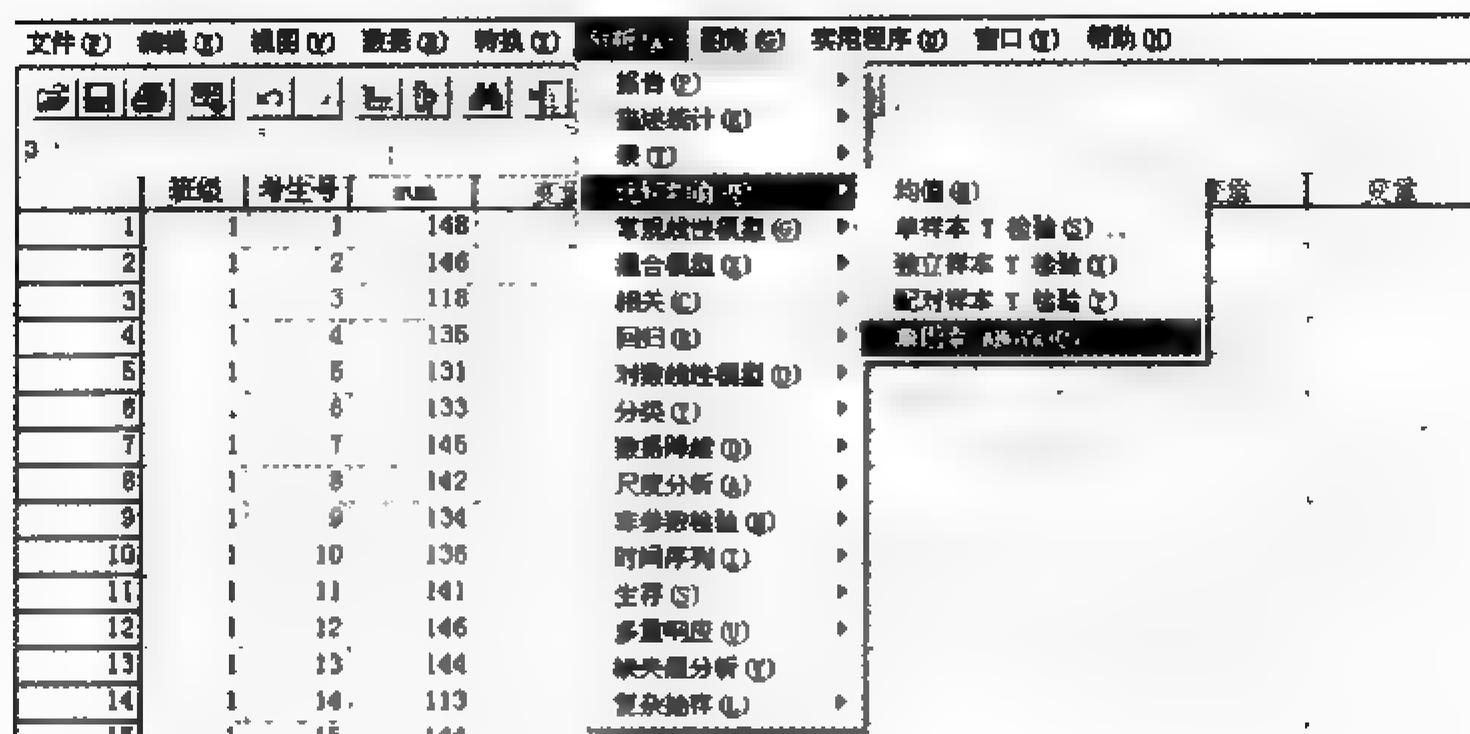


图 4－14

第 3 步，如图 4－15，将左边方框中的“sum”选入右边“因变量列表”栏目下的方框中，将“班级”选入右边“因子”栏目下的方框中。

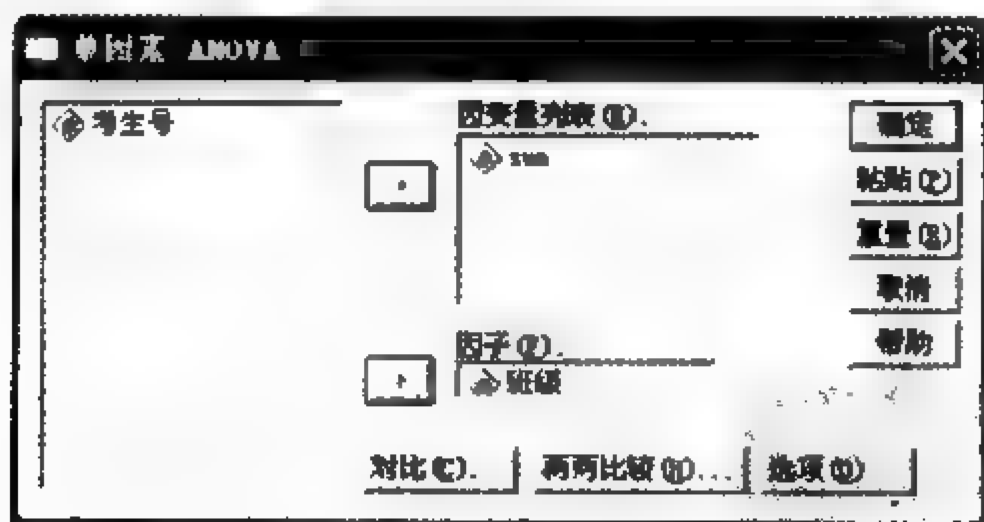


图 4－15

第 4 步，单击“选项”按钮，出现“单因素 ANOVA：选项”子对话框，如图 4－16，在“统计量”栏目下选中“描述性”、“方差同质性检验”，单击“继续”按钮返回图 4－15。单击“两两比较”按钮，出现“单

因素 ANOVA:两两比较”子对话框,如图 4－17,在“假定方差齐性”栏目下选中“LSD(最小显著性差异法)”、“S－N K(多重比较 q 检验)”选项,在“未假定方差齐性”栏目下选中“Tamhane's T2”选项,单击“继续”按钮返回图 4－15。

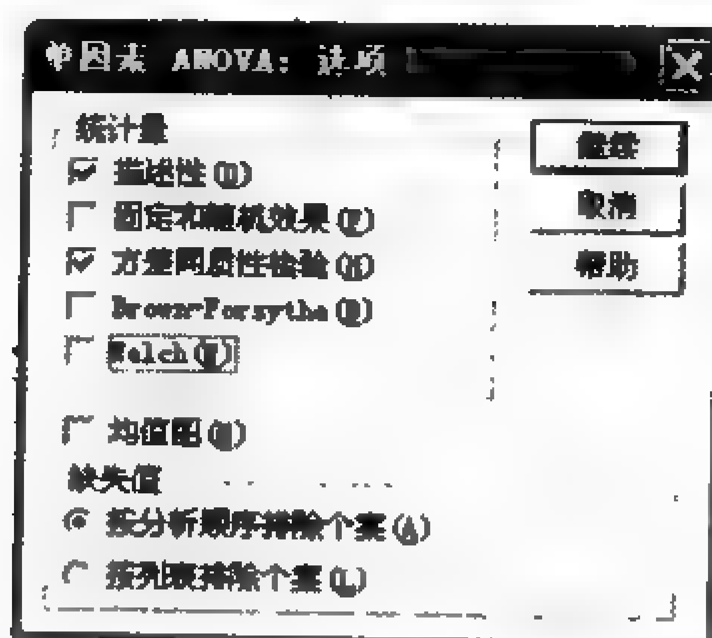


图 4－16

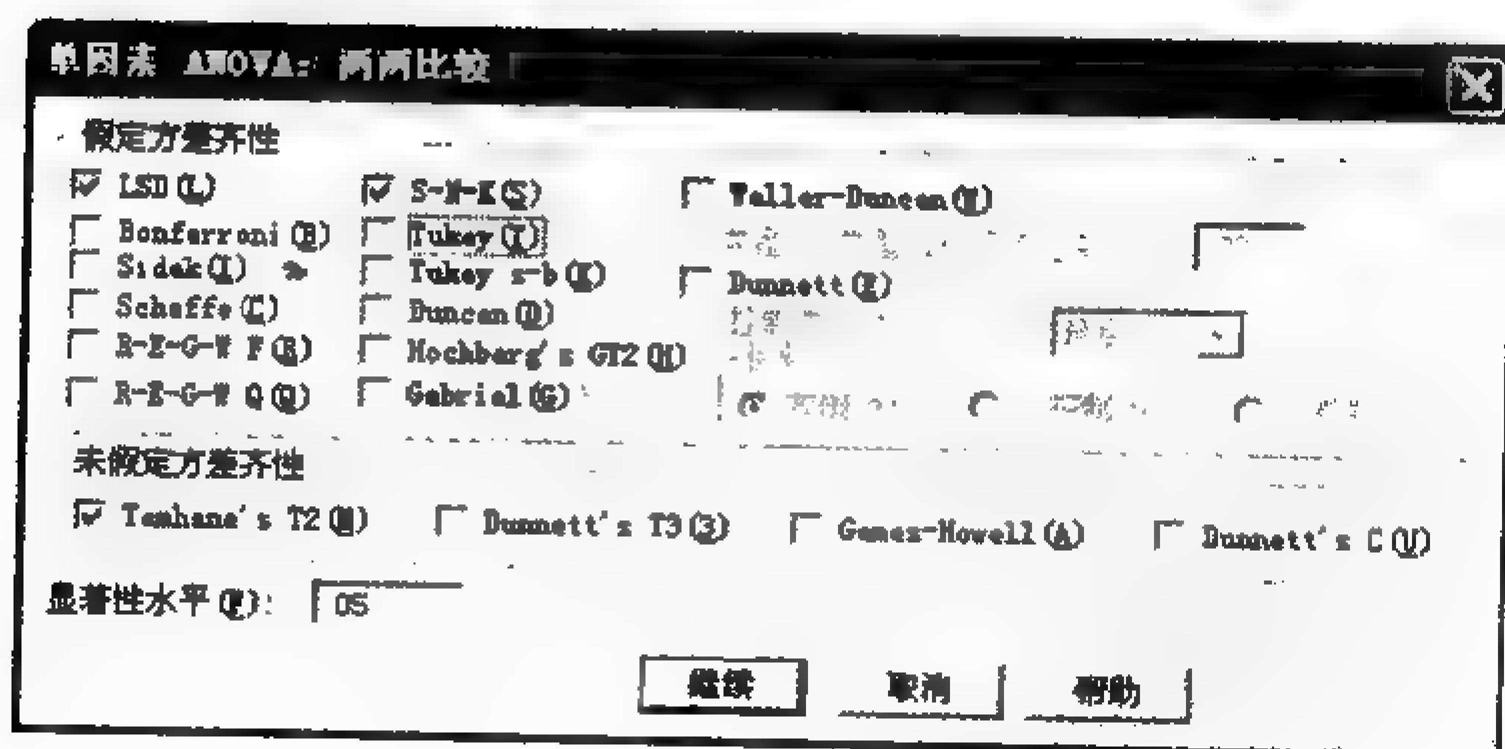


图 4－17

第 5 步,单击图 4－15 中的“确定”按钮,执行单因素方差分析程序。输出结果见表 4－11 至表 4－16。

表 4-11 数学测验成绩统计描述

	人数	均值	标准差	标准误	均值的 95%置信区间		极小值	极大值
					下限	上限		
初三(1)班	50	136.76	7.808	1.104	134.54	138.98	113	148
初三(2)班	49	136.14	8.322	1.189	133.75	138.53	113	146
初三(3)班	50	79.10	44.696	6.321	66.40	91.80	9	148
初三(4)班	50	78.52	42.882	6.064	66.33	90.71	6	143
初三(5)班	50	83.94	42.245	5.974	71.93	95.95	6	144
初三(6)班	50	86.80	39.744	5.621	75.50	98.10	6	143
总　　数	299	100.09	43.243	2.501	95.17	105.01	6	148

表 4-12 测验成绩的方差齐性检验

Levene 统计量	*df*1	*df*2	显著性
43.880	5	293	0.000

根据表 4-12，相伴概率 0.000 小于 0.05，即排出零假设“6 个总体的方差全部相等”，即认为 6 个班所属总体的方差至少有 2 个不相等。

表 4-13 ANOVA(单因素方差分析)

	平方和	*df*	均方	*F*	显著性
组间	198089.642	5	39617.928	32.320	0.000
组内	359154.920	293	1225.785		
总数	557244.562	298			

表 4-13 为单因素方差分析的结果。由于 F 值为 32.320，F 分布的伴随概率为 0.000，即零假设“6 个班的平均成绩相等”成立的概率为 0.000，因此，拒绝零假设，说明 6 个班的平均成绩中至少有两个班不相等。

表4 14 每两个班之间的两两多重比较

	(I)班级	(J)班级	均值差(I-J)	标准误	显著性	95%置信区间	
						下限	上限
LSD	初三(1)班	初三(2)班	.617	7.038	.930	-13.23	14.47
		初三(3)班	57.660(*)	7.002	.000	43.88	71.44
		初三(4)班	58.240(*)	7.002	.000	44.46	72.02
		初三(5)班	52.820(*)	7.002	.000	39.04	66.60
		初三(6)班	49.960(*)	7.002	.000	36.18	63.74
	初三(2)班	初三(1)班	-.617	7.038	.930	-14.47	13.23
		初三(3)班	57.043(*)	7.038	.000	43.19	70.89
		初三(4)班	57.623(*)	7.038	.000	43.77	71.47
		初三(5)班	52.203(*)	7.038	.000	38.35	66.05
		初三(6)班	49.343(*)	7.038	.000	35.49	63.19
	初三(3)班	初三(1)班	-57.660(*)	7.002	.000	-71.44	-43.88
		初三(2)班	-57.043(*)	7.038	.000	-70.89	-43.19
		初三(4)班	.580	7.002	.934	-13.20	14.36
		初三(5)班	-4.840	7.002	.490	-18.62	8.94
		初三(6)班	-7.700	7.002	.272	-21.48	6.08
	初三(4)班	初三(1)班	-58.240(*)	7.002	.000	-72.02	-44.46
		初三(2)班	-57.623(*)	7.038	.000	-71.47	-43.77
		初三(3)班	-.580	7.002	.934	-14.36	13.20
		初三(5)班	-5.420	7.002	.440	-19.20	8.36
		初三(6)班	-8.280	7.002	.238	-22.06	5.50
	初三(5)班	初三(1)班	-52.820(*)	7.002	.000	-66.60	-39.04
		初三(2)班	-52.203(*)	7.038	.000	-66.05	-38.35
		初三(3)班	4.840	7.002	.490	-8.94	18.62

（续表）

	(I)班级	(J)班级	均值差(I J)	标准误	显著性	95%置信区间	
						下限	上限
LSD	初三(5)班	初三(4)班	5.420	7.002	.440	−8.36	19.20
		初三(6)班	2.860	7.002	.683	−16.64	10.92
	初三(6)班	初三(1)班	−49.960(*)	7.002	.000	−63.74	−36.18
		初三(2)班	−49.343(*)	7.038	.000	−63.19	−35.49
		初三(3)班	7.700	7.002	.272	−6.08	21.48
		初三(4)班	8.280	7.002	.238	−5.50	22.06
		初三(5)班	2.860	7.002	.683	−10.92	16.64
Tamhane	初三(1)班	初三(2)班	.617	1.622	1.000	−4.25	5.49
		初三(3)班	57.660(*)	6.417	.000	37.97	77.35
		初三(4)班	58.240(*)	6.164	.000	39.33	77.15
		初三(5)班	52.820(*)	6.076	.000	34.18	71.46
		初三(6)班	49.960(*)	5.728	.000	32.40	67.52
	初三(2)班	初三(1)班	−.617	1.622	1.000	−5.49	4.25
		初三(3)班	57.043(*)	6.432	.000	37.32	76.77
		初三(4)班	57.623(*)	6.180	.000	38.67	76.57
		初三(5)班	52.203(*)	6.092	.000	33.53	70.88
		初三(6)班	49.343(*)	5.745	.000	31.74	66.95
	初三(3)班	初三(1)班	−57.660(*)	6.417	.000	−77.35	−37.97
		初三(2)班	−57.043(*)	6.432	.000	76.77	37.32
		初三(4)班	.580	8.760	1.000	25.71	26.87
		初三(5)班	4.840	8.698	1.000	−30.94	21.26
		初三(6)班	−7.700	8.459	.999	−33.09	17.69

（续表）

	(I)班级	(J)班级	均值差(I-J)	标准误	显著性	95%置信区间	
						下限	上限
Tamhane	初三(4)班	初三(1)班	-58.240(*)	6.164	.000	-77.15	-39.33
		初三(2)班	-57.623(*)	6.180	.000	-76.57	-38.67
		初三(3)班	-.580	8.760	1.000	-26.87	25.71
		初三(5)班	-5.420	8.513	1.000	-30.97	20.13
		初三(6)班	-8.280	8.269	.997	-33.10	16.54
	初三(5)班	初三(1)班	-52.820(*)	6.076	.000	-71.46	-34.18
		初三(2)班	-52.203(*)	6.092	.000	-70.88	-33.53
		初三(3)班	4.840	8.698	1.000	-21.26	30.94
		初三(4)班	5.420	8.513	1.000	-20.13	30.97
		初三(6)班	-2.860	8.203	1.000	-27.48	21.76
	初三(6)班	初三(1)班	-49.960(*)	5.728	.000	-67.52	-32.40
		初三(2)班	-49.343(*)	5.745	.000	-66.95	-31.74
		初三(3)班	7.700	8.459	.999	-17.69	33.09
		初三(4)班	8.280	8.269	.997	-16.54	33.10
		初三(5)班	2.860	8.203	1.000	-21.76	27.48

* 在.05水平上均值差显著。

表4-14显示的是六个班数学平均成绩两两比较的结果，其中上半部分是假设方差齐性下的LSD比较法（最小显著性差异法）多重检验结果，下半部分是假设方差不相等下的Tamhane's法比较法多重检验结果。

根据表4-12我们已经得出6个班级的总体方差不等的结论，在这一前提下，主要看表4-14下半部分的结论，发现初三(1)班、初

三(2)班的平均分与其他4个班的平均分有差异；而它们之间的平均分没有差异。

表4-15进一步给出了Student-Newman-Keuls比较法(即q检验法)多重比较后的均匀分组结果，即将平均分没有差异的总体分在同一个大组中。显然，1班和2班为一组，组内的平均分无差别；3、4、5、6班为一组，组内的平均分无差别；但两个大组之间的平均分存在显著差异。

表4-15 同类子集

	班级	N	alpha=.05 的子集	
			1	2
Student-Newman-Keuls (a, b)	初三(4)班	50	78.52	
	初三(3)班	50	79.10	
	初三(5)班	50	83.94	
	初三(6)班	50	86.80	
	初三(2)班	49		136.14
	初三(1)班	50		136.76
	显著性		.640	.930

将显示同类子集中的组均值。

a 将使用调和均值样本大小=49.831。

b 组大小不相等。将使用组大小的调和均值。将不保证I类错误级别。

三、双因素方差分析

在教育测验中，常常需要研究两种因素对学生学习的影响，例如，人们既想分析学生智力因素对学习效果的影响，也想知道教学方法对学生学习效果的影响。在这种情况下，就要用双因素方差分析方法来处理研究资料。

1. 实施双因素方差分析的前提条件

双因素方差分析的前提条件可以表示成表4-16，其中因素A

分为 k 个水平(或称为方案),因素 B 分为 s 个水平(或称为方案)。所谓"双因素"是指问题中有两个自变量:变量 A 与变量 B,研究这两个变量同时作用于因变量(如:学习效果)。

表 4-16 双因素方差分析的前提条件

		因素 B			
		水平 1	水平 2	…	水平 s
因素 A	水平 1	$x_{111}, x_{112}, \cdots, x_{11n}$	$x_{121}, x_{122}, \cdots, x_{12n}$	…	$x_{1s1}, x_{1s2}, \cdots, x_{1sn}$
	水平 2	$x_{211}, x_{212}, \cdots, x_{21n}$	$x_{221}, x_{222}, \cdots, x_{22n}$	…	$x_{2s1}, x_{2s2}, \cdots, x_{2sn}$
	⋮	⋮	⋮	⋮	⋮
	水平 k	$x_{k11}, x_{k12}, \cdots, x_{k1n}$	$x_{k121}, x_{k22}, \cdots, x_{k2n}$	…	$x_{ks1}, x_{ks2}, \cdots, x_{ksn}$

2. 双因素方差分析的一般步骤

(1) 建立零假设

在双因素方差分析中,总变异 $SS_{总}$ 被分解成四个部分:行间变异(SS_A)、列间变异(SS_B)、交叉变异(SS_{AB})、误差变异(SS_E)共四部分,即

$$SS_{总} = SS_A + SS_B + SS_{AB} + SS_E。 \quad (4.14)$$

因此,相应的零假设包括以下三个部分:

① 假设因素 A 所有水平上的总体平均数相等。即不存在因素 A 效应。

② 假设因素 B 所有水平上的总体平均数相等。即不存在因素 B 效应。

③ 假设因素 A 的总体平均数与因素 B 的总体平均数相等。即不存在因素 A 与 B 的交互效应。

(2) 计算统计量并建立方差分析表

计算 $SS_{总}$、SS_A、SS_B、SS_{AB}、SS_E、MS_A、MS_B、MS_{AB}、MS_E、F 值,并填写在表 4-17 中。

表 4-17 方差分析表

变异来源	平方和	自由度	均方(方差)	F 值	F 值右尾概率 P（显著性）
因素 A	SS_A	$k-1$	MS_A	$F_A = MS_A/MS_E$	
因素 B	SS_B	$s-1$	MS_B	$F_B = MS_B/MS_E$	
因素 A、B 交互效应	SS_{AB}	$(k-1)(s-1)$	MS_{AB}	$F_{AB} = MS_{AB}/MS_E$	
误差因素	SS_E	$ks(n-1)$	MS_E		
总计	$SS_总$	$ksn-1$			

（3）进行统计推断

根据表 4-17，如果 F 值达到 0.05 显著水平的临界值，则说明相应因素的各平均数间差异显著；若 F 值达到 0.01 显著水平的临界值，则说明相应因素的各平均数间差异非常显著。

第四节 EXCEL 与 SPSS 软件应用实例

一、相关系数计算与显著性检验

1. 计算解答题的区分度

【例 4-7】 初三年级共有 439 人，测验卷与例 2-5 中的相同，请分别利用 EXCEL 软件、SPSS 软件用相关系数法计算全年级该次期末数学测验的每道解答题（第 17～25 题）的区分度。

分析：由于解答题的每道题得分与测验总分都是连续型变量，近似地服从正态分布，因此选择使用积差相关法计算。

解法 1：利用 EXCEL 软件中 PEARSON 函数计算。

将区分度放在第 441 行。共分两步进行。

（1）计算第 17 题的区分度。

如图 4－18，在单元格 R441 中键入“=PEARSON(R2:R440，AA2:AA440)”，表示计算考生第 17 题得分与总分的积差相关系数，按 Enter 键，返回值 0.8 就显示在单元格 R441 中。R441 表示的是第 17 题的区分度。

（2）计算第 18～25 题的区分度。

重复第 1 步，即可依次得到第 18～25 题的区分度，计算结果如图 4－18。

R441 =PEARSON(R2:R440, AA2:AA440)

	A	R	S	T	U	V	W	X	Y	Z	AA
1	考生号	x17	x18	x19	x20	x21	x22	x23	x24	x25	sum
2	1	9	9	10	10	12	12	12	14	10	146
3	2	9	9	10	10	1	12	12	11	8	118
4	3	9	9	10	10	12	12	12	14	5	135
5	4	9	9	10	10	12	3	12	14	4	131
438	437	9	9	10	5	12	9	11	13	4	130
439	438	0	0	2	0	0	12	0	0	0	29
440	439	0	0	0	0	0	0	0	0	0	15
441	区分度	0.8	0.8	0.77	0.87	0.77	0.73	0.84	0.88	0.78	

图 4－18

解法 2：利用 SPSS 软件，分三步进行。

（1）如图 4－19，执行【分析】/【相关】/【双变量】程序，出现“双变量相关”对话框。

图 4－19

(2) 如图 4－20，将左边方框中的题目 x1 与总分 sum 选入右边的“变量”下的空框中。在“相关系数”选项下，选中“Pearson”；在“显著性检验”选项下，选中“双侧检验”；选中“标记显著性相关”。

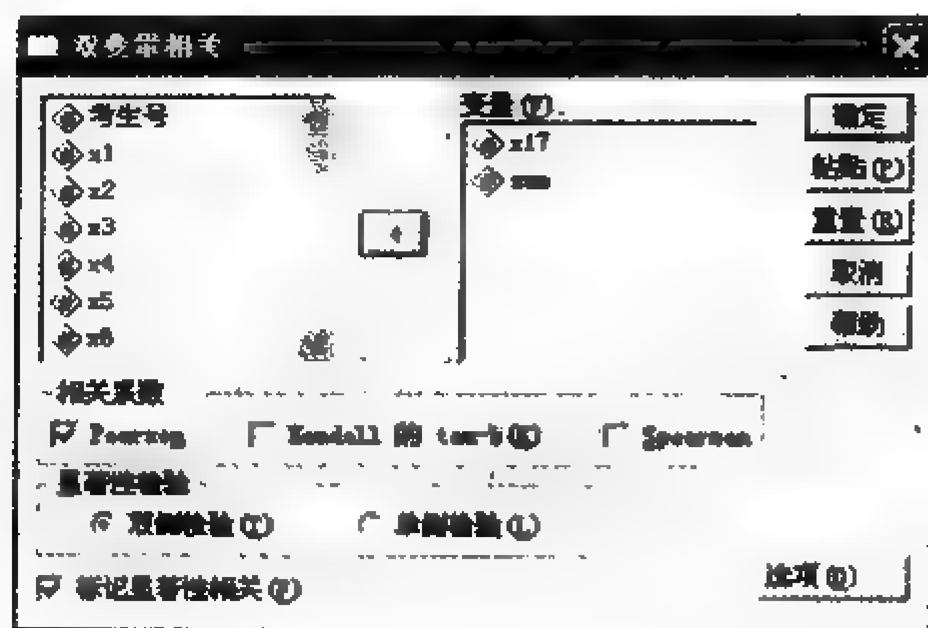


图 4－20

(3) 按“确定”按钮，执行程序计算。输出结果如表 4－18。

表 4－18　第 17 题与测验总分的相关性

		第 17 题	总分
第 17 题	Pearson 相关性 显著性(双侧) N	1 439	.791(* *) .000 439
总分	Pearson 相关性 显著性(双侧) N	.791(* *) .000 439	1 439

* * 在.01 水平(双侧)上显著相关。

表 4－18 显示，样本总数是 439，第 17 题与总分的 Pearson 相关系数值为 0.791，意义水平为 0.01 时，双尾检验的 p 值为 0.000，可以拒绝零假设“第 17 题与总分无关”。

用同样方法，可以依次获得第 18～25 题的试题区分度及其相关显著性水平的检验结果。此处略。

2. 计算选择题与填空题的区分度

【例 4-8】 初二年级共有 439 人，测验卷与例 2-5 中的相同，请利用 EXCEL 软件用相关系数法计算全年级该次期末数学测验的每道选择题与填空题(第 1～16 题)的区分度。

分析：因为选择题与填空题都是二分变量(即只有 3 与 0 两种计分)，而测验总分是连续变量，因此选择使用点二列相关法计算相关系数。

解：将区分度放在第 441 行，以第 1 题的区分度计算为例，共分五步进行。

(1) 计算第 1 题的答对人数与答对率。

如图 4-21，采用函数“COUNTIF”，在单元格 B443 中键入“=COUNTIF(B2:B440,3)”，其意义是“如果单元格中的数据等于 3，则统计个数”，按 Enter 键，返回值 410 就显示在单元格 B443 中。B443 表示的是第 1 题得 3 分的人数。

B443 =COUNTIF(B2:B440,3)

	A	B	C	D	E	F	G	H	I	J	K
439	434	0	3	0	0	0	0	0	0	0	0
440	296	0	0	0	0	0	0	0	0	0	0
441	区分度	0.33									
442								点二列相关系数			
443	答对人数	410									
444	答对率	0.93									
445	答错人数	29									
446	答错率	0.07									
447	答对平均数	96.2									
448	答错平均数	40.1									
449	总分标准差	42.9									

图 4-21

在单元格 B444 中键入“=B440/439”，其意义是“得 3 分的人数与总人数的比”，按 Enter 键，返回值 0.93 就显示在单元格 B444 中。B444 表示的是第 1 题得 3 分的比率，即答对率。

(2) 计算第 1 题的答错人数与答错率。

如图 4-22，在单元格 B445 中键入“=439−B443”，其意义是“总人数减去答对人数”，按 Enter 键，返回值 29 就显示在单元格

B445 中。B445 表示的是第 1 题得 0 分的人数。也可以仿照第 1 步的方法，用“＝COUNTIF(B2:B440,0)”计算，有兴趣的读者可以试试。

B445	=439-B443										
	A	B	C	D	E	F	G	H	I	J	K
439	434	0	3	0	0	0	0	0	0	0	0
440	296	0	0	0	0	0	0	0	0	0	0
441	区分度	0.33									
442								点二列相关系数			
443	答对人数	410									
444	答对率	0.93									
445	答错人数	29									
446	答错率	0.07									
447	答对平均数	96.2									
448	答错平均数	40.1									
449	总分标准差	42.9									

图 4－22

在单元格 B446 中键入“＝B445/439”，其意义是“得 0 分的人数与总人数的比”，按 Enter 键，返回值 0.07 就显示在单元格 B446 中。B446 表示的是第 1 题得 0 分的比率，即答错率。

（3）计算第 1 题的答对平均数。

答对平均数的意义是答对第 1 题的所有考生测验总分的平均分。

先用鼠标选中区域 B1:AA440，即将所有考生的数据选中，然后执行【数据】/【排序】程序，出现“排序”对话框，如图 4－23。在“主要关键字”下空框中选中“x1”（即第 1 题），方框右边选择降序，按“确定”对话框，返回到 EXCEL 工作表中。这时，表中的数据按照 x1（第 1 题）这列数据排序，其余数据的次序相应随之重排。这时，“sum”（即总分）列的前 410 个数据就是答对第 1 题的所有考生的测验总分。



图 4－23

如图4 24,采用函数“AVERAGE”计算答对第1题的所有考生的测验总分,在单元格B447中键入“=AVERAGE(AA2:AA411)”,按Enter键,返回值96.2就显示在单元格B447中。B447表示的是第1题答对平均数。

B447 =AVERAGE(AA2:AA411)

	A	B	C	D	E	F	G	H	I	J	K
439	434	0	3	0	0	0	0	0	0	0	0
440	296	0	0	0	0	0	0	0	0	0	0
441	区分度	0.33									
442								点二列相关系数			
443	答对人数	410									
444	答对率	0.93									
445	答错人数	29									
446	答错率	0.07									
447	答对平均数	96.2									
448	答错平均数	40.1									
449	总分标准差	42.9									

图4-24

同样地,采用函数“AVERAGE”计算答错第1题的所有考生的测验总分,在单元格B448中键入“=AVERAGE(AA412:AA440)”,按Enter键,返回值40.1就显示在单元格B448中。B448表示的是第1题答错平均数,如图4-25。

B448 =AVERAGE(AA412:AA440)

	A	B	C	D	E	F	G	H	I	J	K
439	434	0	3	0	0	0	0	0	0	0	0
440	296	0	0	0	0	0	0	0	0	0	0
441	区分度	0.33									
442								点二列相关系数			
443	答对人数	410									
444	答对率	0.93									
445	答错人数	29									
446	答错率	0.07									
447	答对平均数	96.2									
448	答错平均数	40.1									
449	总分标准差	42.9									

图4-25

(4)计算所有考生总分的标准差。

采用函数“STDEV”计算,在单元格B449中键入“=STDEV(AA2:AA440)”,按Enter键,返回值42.9就显示在单元格B449

中。B449 表示的是测验卷所有考生总分的标准差。

(5) 计算第 1 题的区分度。

如图 4-26,在单元格 B441 中键入“ SQRT(B444 * B446) * (B447-B448)/B449”,即公式 $r_{pb}=\frac{\bar{x}_p-\bar{x}_q}{\sigma_t}\sqrt{pq}$,按 Enter 键,返回值 0.33 就显示在单元格 B441 中。B441 表示的是测验卷第 1 题的区分度。

B441　=SQRT(B444*B446)*(B447-B448)/B449

	A	B	C	D	E	F	G	H	I	J	K	L
439	434	0	3	0	0	0	0	0	0	0	0	0
440	296	0	0	0	0	0	0	0	0	0	0	0
441	区分度	0.33										
442							点二列相关系数					
443	答对人数	410										
444	答对率	0.93										
445	答错人数	29										
446	答错率	0.07										
447	答对平均数	96.2										
448	答错平均数	90.1										
449	总分标准差	42.9										

图 4-26

二、回归方程计算与有效性检验

1. 一元线性回归问题

【例 4-9】 用 SPSS 软件计算例 4-4 中两次测验成绩的一元线性回归方程,并进行有效性检验。

解:回归方程的计算分为四步进行。

(1) 如图 4-27,将表 4-7 中的数据导入 SPSS 数据编辑器的工作表中,其中 x 表示上学期期末数学测验成绩,y 表示下学

文件(F)　编辑(E)　视图(V)　数据(D)　转换(T)　分析(A)　图形(G)　实用程序(U)　窗口(W)　帮助(H)

	学号	x	y	变量
1	1	80	76	
2	2	60	66	
3	3	82	80	
4	4	72	75	
5	5	73	69	
6	6	80	85	
7	7	86	90	
8	8	90	84	
9	9	85	93	
10	10	78	78	
11	11	95	93	

数据视图　变量视图　　SPSS 处理

图 4-27

期期末数学测验成绩。

(2) 如图 4－28，执行【分析】/【回归】/【线性】程序，出现“线性回归”对话框。



图 4－28

(3) 如图 4－29，将左边方框中的“下学期期末数学成绩(y)”选入右边“因变量”栏目下的方框中，将“上学期期末数学成绩(x)”选入右边“自变量”栏目下的方框中。单击“统计量”按钮，出现“线性回归：统计量”子对话框，如图 4－30。



图 4－29

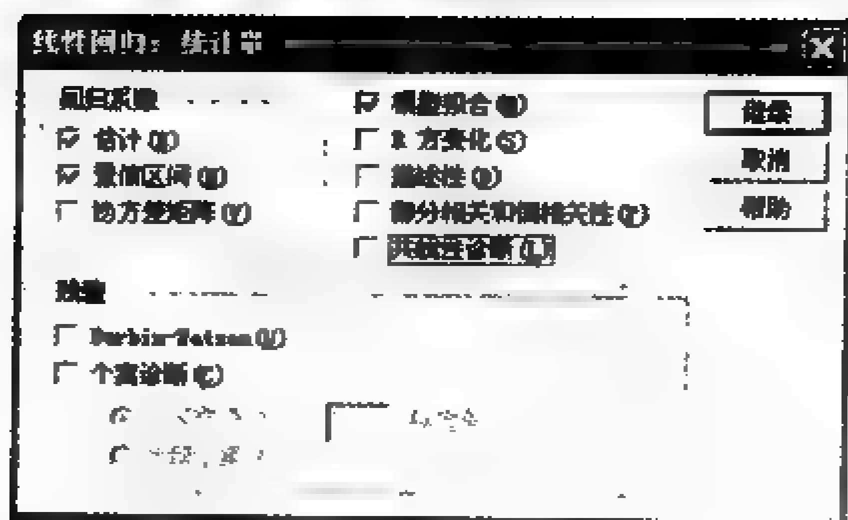


图 4－30

（4）在“线性回归：统计量”子对话框中，在“回归系数”栏目下选中“估计”、“置信区间”、“模型拟合”，单击“继续”按钮，返回图4－29。

根据需要，可以进一步定义图 4－29 中的子对话框“图”、“保存”、“选项”。本例选择系统默认。最后，单击“确定”按钮，执行线性回归分析程序。输出结果见表 4－19 至表 4－21。

表 4－19　模型摘要

模型	R	R^2	调整的 R^2	估计的标准差
1	.908(a)	.824	.815	4.691

a　预测变量：(常量)，上学期期末数学成绩。

表 4－20　方差分析表(b)

模型		平方和	自由度	均方(方差)	F 值	显著性
1	回归	1860.091	1	1860.091	84.526	.000(a)
	残差	396.109	18	22.006		
	合计	2256.200	19			

a　预测变量：(常量)，上学期期末数学成绩。
b　因变量：下学期期末数学成绩。

表 4－19 中，第 1 列说明回归分析采用的是模型 1；第 2 列两个变量之间的相关系数 R 值是 0.908，说明两个变量的相关程度很高；第 3 列确定系数 R^2 值是 0.824，说明常数和自变量“上学期期末数

学成绩”可以解释因变量“下学期期末数学成绩”中的82.4%；第4列调整的R^2值是0.815；第5列估计的标准差值为4.691，表示的是根据回归方程用自变量估计因变量的残差标准差。

在表4-20中，用方差分析法检验回归方程的有效性。由于F值为84.526，F分布的伴随概率为0.000，即零假设“回归系数为0”成立的概率为0.000，因此，拒绝零假设，说明回归方程有效。即上、下学期期末数学成绩之间存在线性相关关系。

表4-21　回归系数(a)

模型		非标准化系数		标准化系数	T值	显著性	B的95%置信区间	
		B	标准误	Beta			下限	上限
1	(常量)	12.575	7.010		1.794	0.090	-2.153	27.303
	上学期期末数学成绩	0.847	0.092	0.908	9.194	0.000	0.653	1.040

a　因变量：下学期期末数学成绩。

在表4-21中，首先交代了线性回归模型是1，估计参数包括常量a和回归系数b（即自变量x的系数）。在没有进行标准化处理前，求得常量a的估计值是12.575，标准误为7.010；回归系数b的估计值是0.847，标准误是0.092。由于数据属于小样本，因此服从t分布，计算常量a和回归系数b相应的t值分别为1.794、9.194。回归系数b的伴随概率是0.000，表示用t统计量检验零假设“回归系数为0”的概率是0.000，因此，拒绝零假设，同样说明上、下学期期末数学成绩之间存在线性相关关系。

所以，所求得的两次测验成绩之间一元线性回归方程为$\hat{y}=0.847x+12.575$，这个结论与例4-4中的相同。

2. 多元线性回归问题

【例4-10】　表4-22是某学校20名高三学生调研测试、一模、二模的数学成绩，以及高考数学成绩，其中x_1表示调研测试成绩，x_2

表示一模成绩，x_3 表示二模成绩，y 表示高考成绩。试用 SPSS 软件计算四次测验成绩的多元线性回归方程，并进行有效性检验，为下一届高三学生的高考成绩预测做准备。

表 4-22　20 名高三学生四次数学测验成绩

序号	1	2	3	4	5	6	7	8	9	10
x_1	142	136	137	118	98	84	133	104	52	112
x_2	136	120	140	109	89	96	145	115	64	107
x_3	140	128	138	120	107	98	140	106	73	120
y	130	135	142	122	102	106	138	113	75	115

序号	11	12	13	14	15	16	17	18	19	20
x_1	87	99	40	107	76	121	66	122	116	79
x_2	79	82	50	101	88	124	49	103	114	94
x_3	92	108	65	133	90	118	60	116	104	93
y	103	100	70	111	94	130	72	127	117	82

解：回归方程的计算分为四步进行。

（1）如图 4-31，将表 4-22 中的数据导入 SPSS 数据编辑器的工作表中。

	学号	x1	x2	x3	y
1	1	142	136	140	130
2	2	136	120	128	135
3	3	137	140	138	142
4	4	118	109	120	122
5	5	98	89	107	102
6	6	84	96	98	106
7	7	133	145	140	138
8	8	104	115	106	113
9	9	52	64	73	75
10	10	112	107	120	115
11	11	87	79	92	103
12	12	99	82	108	100

图 4-31

(2) 如图 4－32，执行【分析】/【回归】/【线性】程序，出现“线性回归”对话框。

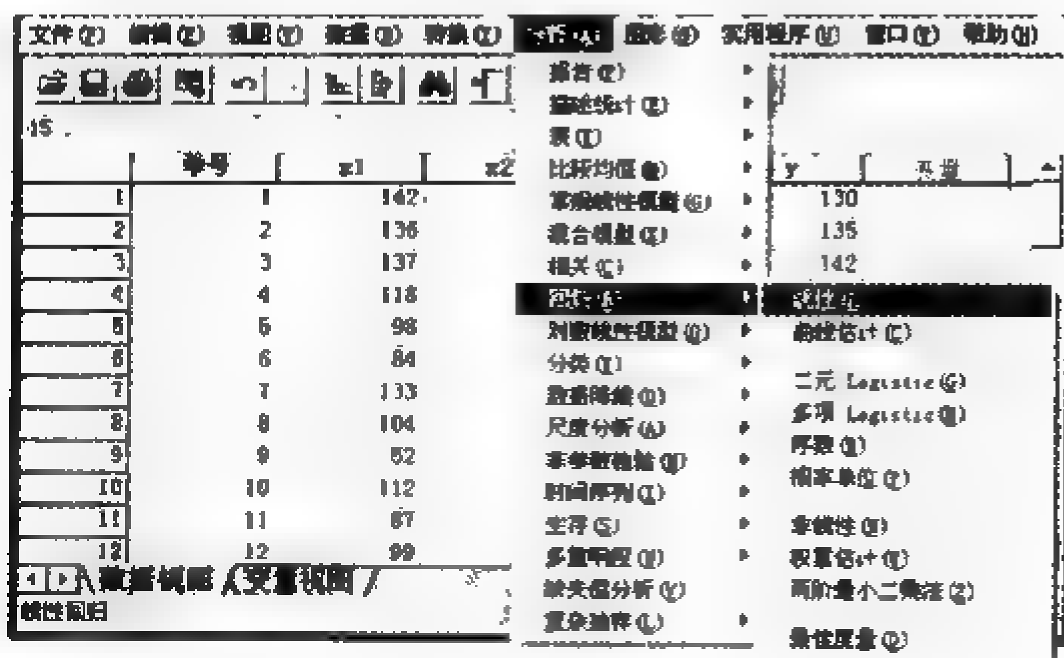


图 4－32

(3) 如图 4－33，将左边方框中的“高考成绩[y]”选入右边“因变量”栏目下的方框中，将“调研测试成绩[x1]、一模成绩[x2]、二模成绩[x2]”选入右边“自变量”栏目下的方框中。依次单击“统计量”、“图”、“保存”、“选项”按钮，设定方法参见例 4－9。

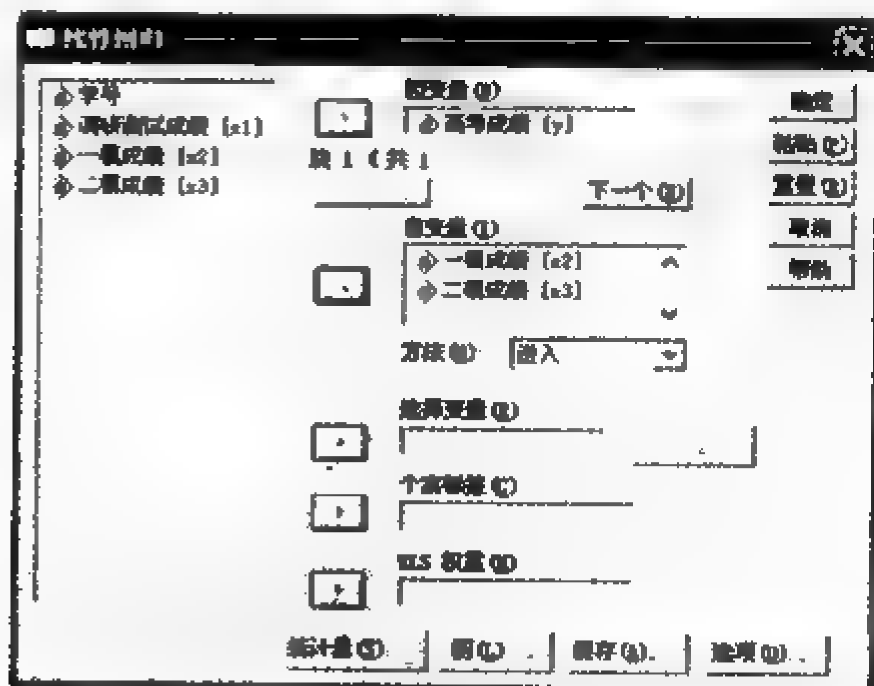


图 4－33

(4) 单击“确定”按钮，执行线性回归分析程序。输出结果见表

4 -23 至表4 25。

表 4 -23 模型摘要(b)

模型	R	R 方	调整的 R 方	估计的标准差
1	0.969(a)	0.939	0.928	5.945

a 预测变量:(常量),二模成绩,一模成绩,调研测试成绩。
b 因变量:高考成绩。

表 4 23 中,第 1 列说明回归分析采用的是模型 1;第 2 列 4 个变量之间的复相关系数 R 值是 0.969,说明自变量"二模成绩、一模成绩、调研测试成绩"与因变量"高考成绩"的相关程度很高;第 3 列确定系数 R^2 值是 0.939,说明自变量"二模成绩、一模成绩、调研测试成绩"与常数可以解释因变量"高考成绩"中的 93.9%;第 4 列调整的 R^2 值是 0.928;第 5 列估计的标准差值为 5.945,表示的是根据回归方程用自变量估计因变量的残差标准差。

表 4-24 方差分析表(b)

模型		平方和	自由度	均方(方差)	F 值	显著性
1	回归	8709.673	3	2903.224	82.139	0.000(a)
	残差	565.527	16	35.345		
	合计	9275.200	19			

a 预测变量:(常量),二模成绩,一模成绩,调研测试成绩。
b 因变量:高考成绩。

在表 4-24 中,用方差分析法检验多元线性回归方程的有效性。由于 F 值为 82.139,F 分布的伴随概率为 0.000,即零假设"全部回归系数为 0"成立的概率为 0.000,因此,拒绝零假设,说明回归方程有效。即二模成绩、一模成绩、调研测试成绩与高考成绩之间存在线性相关关系。

在表 4 25 中,首先交代了线性回归模型是 1,估计参数包括常量 a 和回归系数 b_1、b_2、b_3(即自变量调研测试成绩、一模成绩、二模成绩的系数)。在没有进行标准化处理前,求得常量 a 的估计值是

30.229，标准误为 6.836；回归系数 b_1 的估计值是 0.506，标准误是 0.147；回归系数 b_2 的估计值是 0.240，标准误是 0.137；回归系数 b_3 的估计值是 0.033，标准误是 0.173。由于数据属于小样本，因此服从 t 分布。计算常量 a 相应的 t 值为 4.422，相应的伴随概率是 0.000，显然，用 t 统计量检验零假设 1“常数 a 为 0”的概率是 0.000，因此，拒绝零假设 1。计算回归系数 b_1 相应的 t 值为 3.439，相应的伴随概率是 0.003，显然，用 t 统计量检验零假设 2“回归系数 b_1 为 0”的概率是 0.003，因此，拒绝零假设 2，认为调研测试成绩与高考成绩之间存在线性相关关系。计算回归系数 b_2 相应的 t 值为 1.755，相应的伴随概率是 0.098，显然，用 t 统计量检验零假设 2“回归系数 b_2 为 0”的概率是 0.098，可以拒绝零假设 2，认为一模成绩与高考成绩之间存在线性相关关系。计算回归系数 b_3 相应的 t 值为 0.193，相应的伴随概率是 0.849，这时用 t 统计量检验零假设 2“回归系数 b_3 为 0”的概率是 0.849，这时应接受零假设 2，认为二模成绩与高考成绩之间不存在线性相关关系。

表 4－25　回归系数(a)

模型		非标准化系数		标准化系数	T 值	显著性	B 的 95% 置信区间	
		B	标准误	Beta			下限	上限
1	(常量)	30.229	6.836		4.422	0.000	15.737	44.720
	调研测试成绩	0.506	0.147	0.658	3.439	0.003	0.194	0.818
	一模成绩	0.240	0.137	0.294	1.755	0.098	－0.050	0.529
	二模成绩	0.033	0.173	0.036	0.193	0.849	0.334	0.401

a　因变量：高考成绩。

这时，从自变量中剔除“二模成绩”，再重复上述 1—4 步，执行多元线性回归分析后，可以得到表 4－26、表 4－27、表 4－28。

表 4-26　模型摘要(b)

模型	R	R^2	调整的 R^2	估计的标准差
1	0.969(a)	0.939	0.932	5.774

a　预测变量:(常量),一模成绩,调研测试成绩。
b　因变量:高考成绩。

表 4-27　方差分析表(b)

模型		平方和	自由度	均方(方差)	F 值	显著性
1	回归	8708.356	2	4354.178	130.585	0.000(a)
	残差	566.844	17	33.344		
	合计	9275.200	19			

a　预测变量:(常量),一模成绩,调研测试成绩。
b　因变量:高考成绩。

表 4-28　回归系数(a)

模型		非标准化系数		标准化系数	T 值	显著性	B 的 95% 置信区间	
		B	标准误	Beta			下限	上限
1	(常量)	31.076	5.091		6.104	0.000	20.334	41.818
	调研测试成绩	0.523	0.115	0.680	4.548	0.000	0.280	0.765
	一模成绩	0.250	0.122	0.307	2.050	0.056	−0.007	0.508

a　因变量:高考成绩。

对表 4-26、表 4-28、表 4-28 的解释参见前面说明。

所以,所求得的测验成绩间多元线性回归方程为 $\hat{y}=0.523x_1+0.25x_2+31.076$。

三、方差分析与平均数差异检验

1. 运用单因素方差分析进行班级间成绩对比研究

【例 4-11】　在例 4-6 中,已经使用 SPSS 软件,运用单因素方

差分析法分析了某校初三年级六个班上学期期末测验数学成绩的差异情况，现在请用 EXCEL 软件对该例中六个班的数学测验成绩进行分析与比较。

解：用 EXCEL 软件共分 5 步完成。

(1) 打开已经输入数据的 EXCEL 工作表，将数据整理成图 4-34 的排列形式。

6个班的数学成绩

1班	2班	3班	4班	5班	6班
148	119	141	79	24	143
146	146	137	42	112	99
118	146	18	107	62	102
135	142	18	27	45	90
131	138	118	80	42	123
133	140	9	118	87	73
145	136	98	106	13	56
142	132	84	58	12	65
134	126	133	13	103	118

图 4-34

(2) 如图 4-35，执行【工具】/【数据分析】程序，出现“数据分析”对话框。

图 4　35

（注：如果“工具”菜单中没有出现“数据分析”命令，需要先按照“分析工具库”，安装方法为：在“工具”菜单中单击“加载宏”命令，选中“分析工具库”复选框即可）

（3）如图 4－36，在“数据分析”对话框的“分析工具”栏目下选中“方差分析：单因素方差分析”选项，然后单击“确定”按钮，出现“方差分析：单因素方差分析”对话框。



图 4－36

（4）如图 4－37，在“方差分析：单因素方差分析”对话框中，“输入区域”栏目后的方框中输入 \$A\$2：\$F\$52（即用鼠标选中 A2：F52 的数字区域），“分组方式”栏目选中“列”，选中“标志位于第一行”；显著性水平为“0.05”；“输出选项”栏目下选中“输出区域”，在工作表中选择一个单元格（如 \$I\$4，即 I4）作为输出区域左上角的第一个单元格。

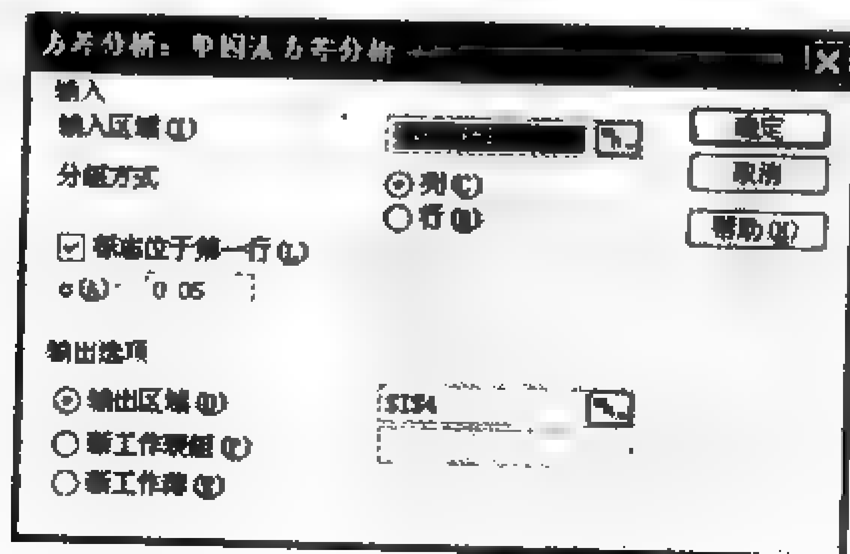


图 4－37

(5) 单击“确定”按钮，则在工作表的输出区域内显示单因素方差分析“计算表”和“分析表”，如图 4 - 38。

	A
2	1班
3	148
4	146
5	118
6	135
7	131
8	133
9	145
10	142
11	134
12	136
13	141
14	146
15	144
16	113
17	144
18	143
19	136
20	137
21	135

方差分析：单因素方差分析

SUMMARY

组	观测数	求和	平均	方差
1班	50	6838	136.76	60.96163
2班	50	6721	134.42	216.2486
3班	50	3955	79.1	1997.765
4班	50	3926	78.52	1838.867
5班	50	4197	83.94	1784.67
6班	50	4340	86.8	1579.592

方差分析

差异源	SS	df	MS	F	P-value	F crit
组间	193318.14	5	38663.63	31.02147	2.61E-25	2.2447033
组内	366427.1	294	1246.351			
总计	559745.24	299				

例4-12

图 4 - 38

根据图 4 - 38 中的数据作出统计推断。

由于 $F = 31.02 > 2.24 = F_{0.05}$，根据右尾 F 检验推断规则，在 0.05 显著性水平上拒绝零假设“六个班平均分无差异”，因此，作出推断：六个班的平均分有显著差异。

从六个班的平均分可以看出，1 班和 2 班作为一个整体，与其他 4 个班作为一个整体，两个部分之间差异明显。至于 4 班至 6 班是否有明显差异，可以用函数“TTEST”做进一步的检验。（此处略）

2. 运用双因素方差分析测验卷的等值问题与学生的学习差异

【例 4 - 12】 A1、A2、A3、A4 等四所中学的高三学生先后做三套高考数学模拟测验卷，测验平均分如表 4 - 29，请用 SPSS 软件分析三套高考数学模拟测验卷是否等值？四所学校的学生水平是否一致？

表 4-29　四所中学三套测验卷平均分

		模拟测验卷		
		B1	B2	B3
学校	A1	97.5	125	112
	A2	88.5	118	108
	A3	81	120	105
	A4	78	115	110

解：测验卷的等值检验与学校学生水平检验的方差分析分为四步进行。

(1) 如图 4-39，将表 4-29 中的数据导入 SPSS 数据编辑器的工作表中，注意导入时的数据排列方式。

	学校	测验卷	数学平均分
1	A1	B1	97.50
2	A1	B2	125.00
3	A1	B3	112.00
4	A2	B1	88.50
5	A2	B2	118.00
6	A2	B3	108.00
7	A3	B1	81.00
8	A3	B2	120.00
9	A3	B3	105.00
10	A4	B1	78.00
11	A4	B2	115.00
12	A4	B3	110.00

图 4-39

(2) 如图 4-40，执行【分析】/【常规线性模型】/【单变量】程序，出现“单变量”对话框(特别说明：这里的单变量指的是单因变量)，如图 4-41。



图 4-40

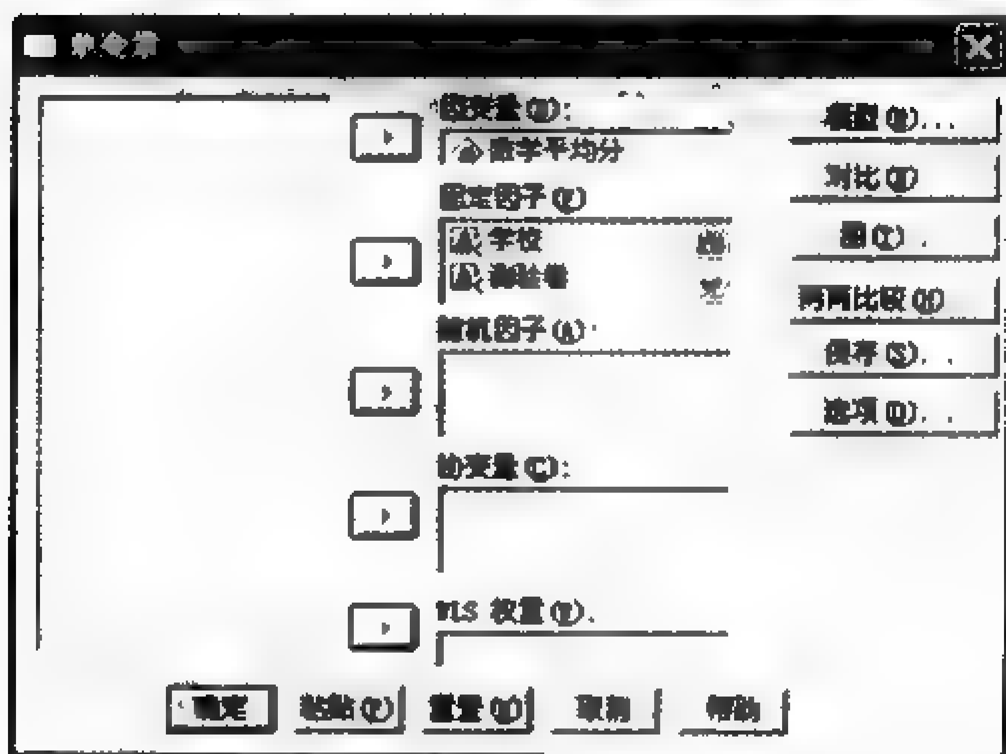


图 4-41

(3) 在图 4-41 中，把左边“数学平均分”变量导入到右边“因变量”下的空框中，把左边“学校”、“测验卷”变量导入到右边“固定因子”下的空框中。

(4) 对“单变量”中的六个中的五个按钮“模型”、“对比”、“图”、“两两比较”、“选项”中的选项进行选择，基本的选择项与选择方法如图 4-42、4-43、4 44、4-45、4 -46 所示(“保存”选项没有操作，有兴趣的读者可以自行探索不同的选择，再将分析结果进行比

较)，每个选项选择完后，单击“继续”按钮，返回图 4 41。

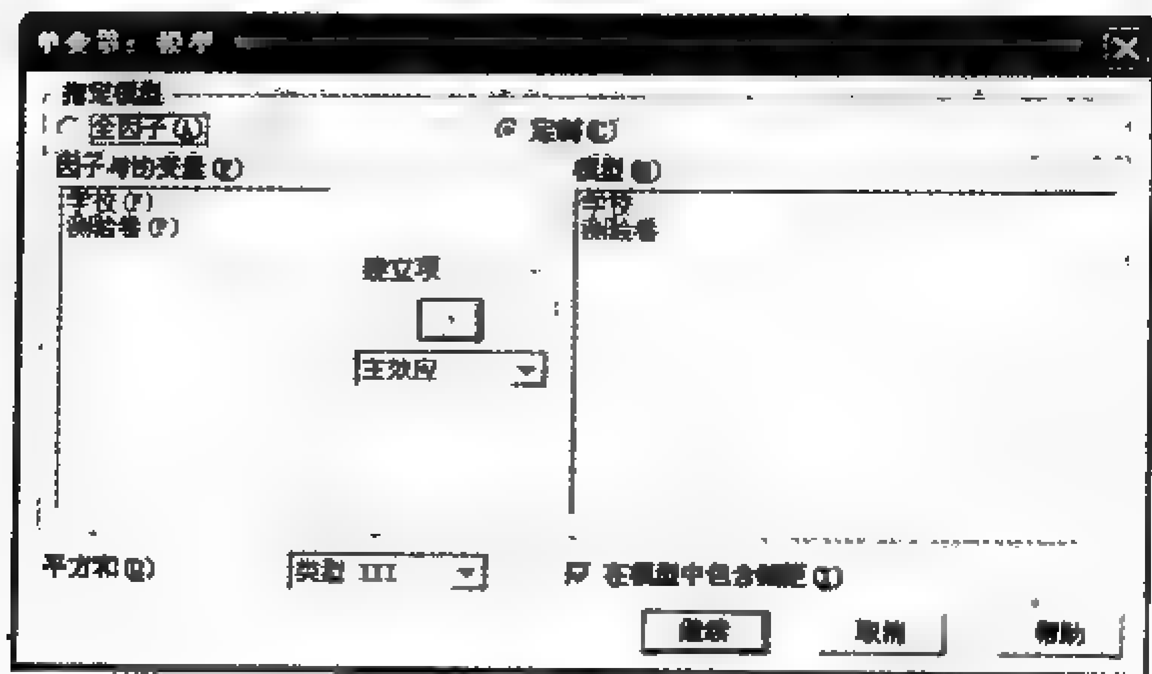


图 4－42

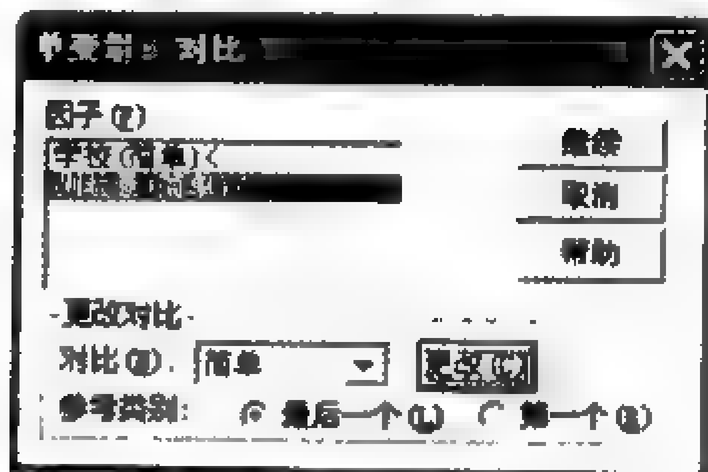


图 4－43

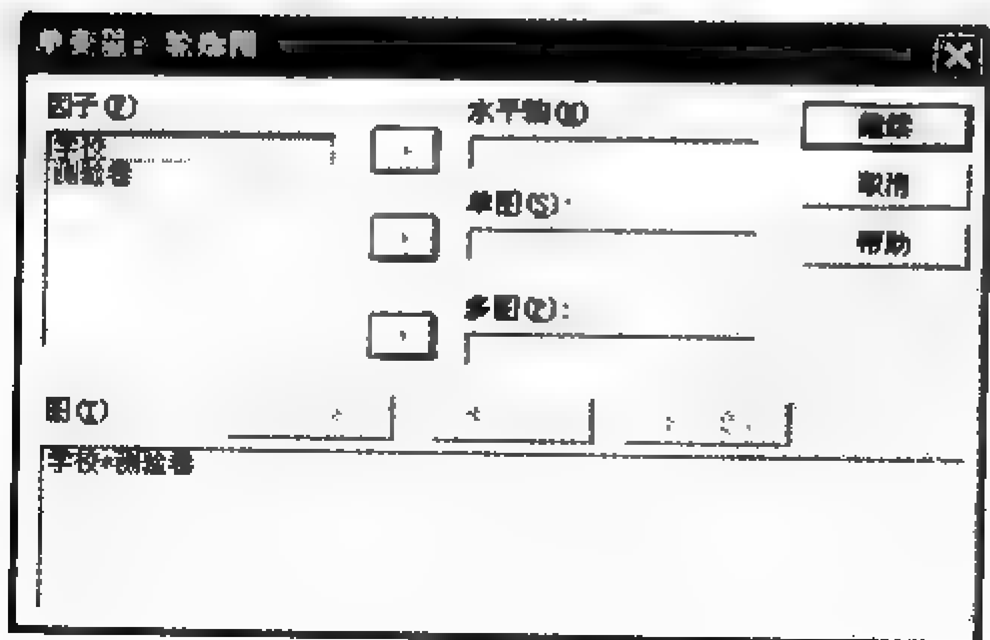


图 4－44

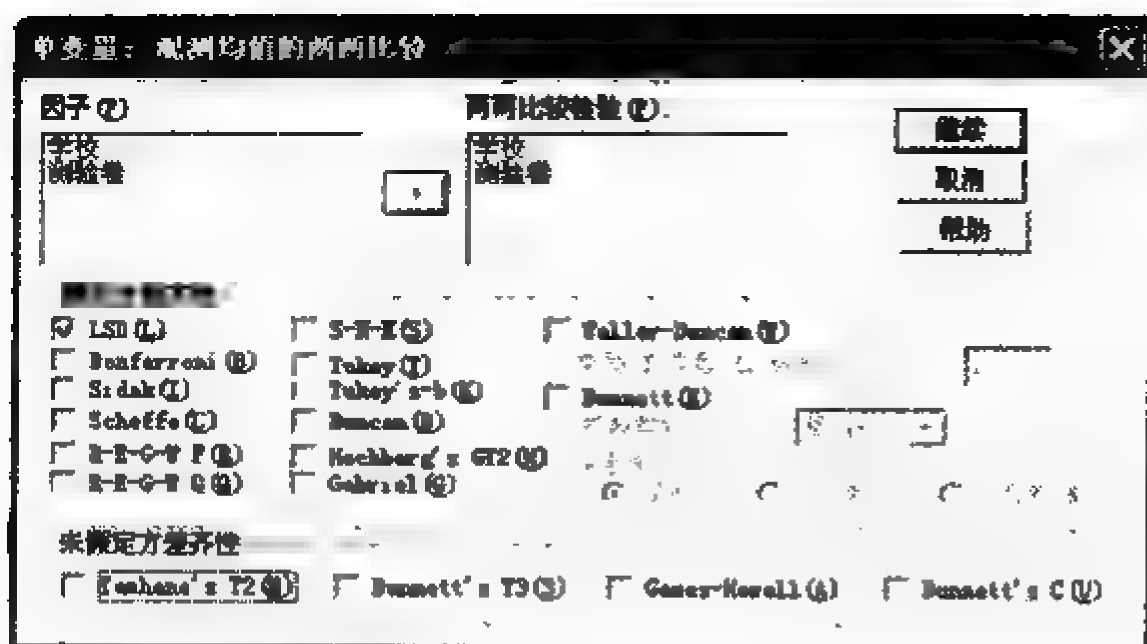


图 4－45

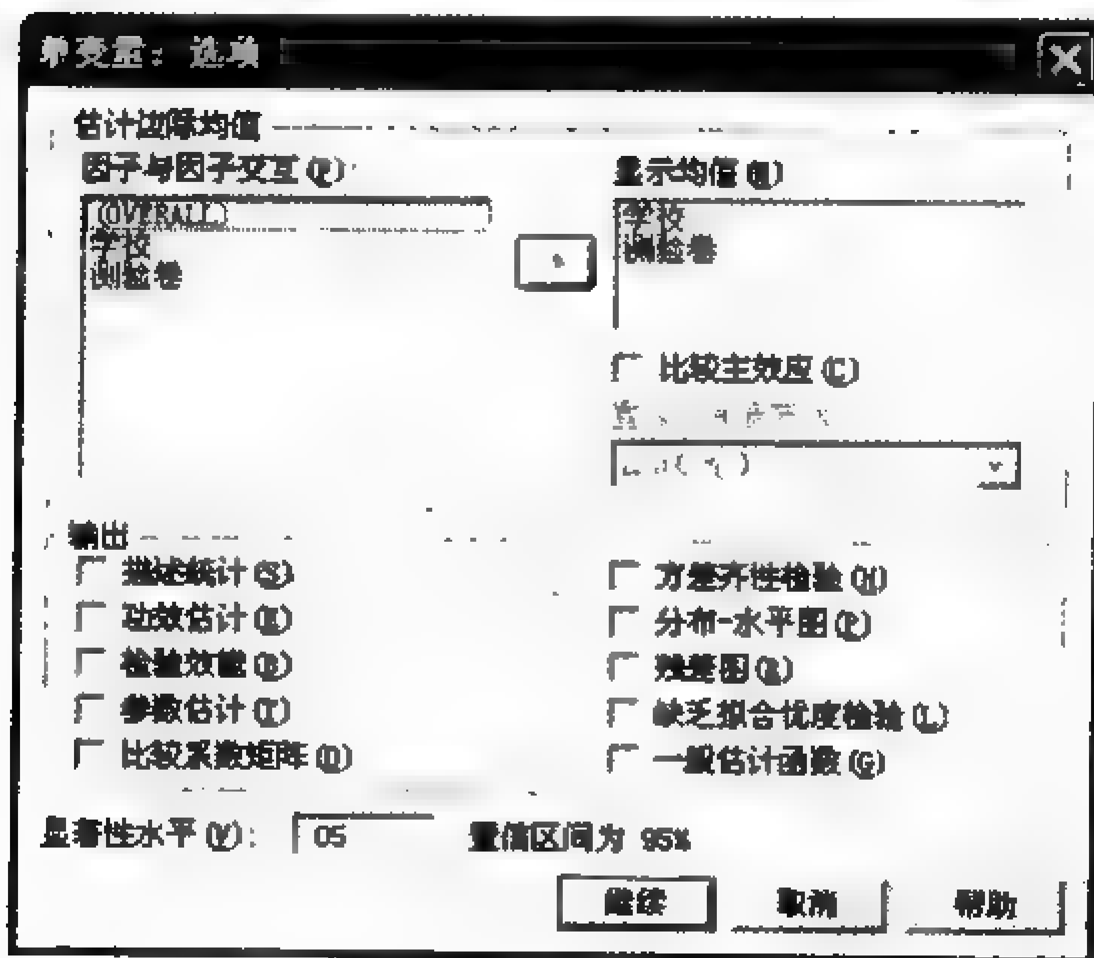


图 4－46

（5）单击图 4－41 中的“确定”按钮，执行单因变量双因素方差分析程序。输出结果见表 4－30 至表 4－38。

表 4-30　主体间效应的检验

因变量:数学平均分

变差来源	Ⅲ型平方和	自由度 df	均方	F 值	显著性 Sig.
校正模型	2504.667(a)	5	500.933	28.489	.000
截距	131880.333	1	131880.333	7500.303	.000
学校	201.500	3	67.167	3.820	.076
测验卷	2303.167	2	1151.583	65.493	.000
误差	105.500	6	17.583		
总计	134490.500	12			
校正的总计	2610.167	11			

a　回归系数平方=0.960(调整后的回归系数平方=0.926)。

表 4-30 中,第一列是变差的来源,其中"校正模型的变差"=学校的变差+测验卷的变差,这里校正去掉了截距项。"校正的总计"=校正模型的变差+误差项的变差。第六列的相伴概率值均小于 0.01,说明在学校与测验卷的不同水平、不同组合中,至少有的效果之间,有显著性差异。

表 4-31　不同学校数学平均分的对比结果

学　　校			因变量
简单对比(a)			数学平均分
级别 1 和级别 4	对比估算值		10.500
	假设值		0
	差分(估计-假设)		10.500
	标准误		3.424
	Sig.		.022
	差分的 95%置信区间	下限	2.122
		上限	18.878

（续表）

学　校			因变量
简单对比(a)			数学平均分
级别 2 和级别 4	对比估算值		3.833
	假设值		0
	差分(估计－假设)		3.833
	标准误		3.424
	Sig.		.306
	差分的 95%置信区间	下限	−4.544
		上限	12.211
级别 3 和级别 4	对比估算值		1.000
	假设值		0
	差分(估计－假设)		1.000
	标准误		3.424
	Sig.		.780
	差分的 95%置信区间	下限	−7.378
		上限	9.378

a．参考类别＝4。

表 4－31 给出了四所学校的数学平均分的比较结果，由于级别 1(学校 A1)与级别 4(学校 A4)的相伴概率值为 0.022＜0.05，表示学校 A1 与学校 A4 的数学平均分有显著性差异；同样方法，我们可以看出学校 A2 与学校 A4、学校 A3 与学校 A4 间的数学平均分并没有显著性差异。

表 4－32　不同学校之间数学平均分的检验结果

源	平方和	*df*	均方	*F*	Sig.
对比	201.500	3	67.167	3.820	.076
误差	105.500	6	17.583		

表 4－32 表明，从方差分析的角度看，由于相伴概率值 0.076＞0.05，因此，不同学校之间的数学平均分的差异不显著。

表 4-33　不同数学测验卷之间平均分的检验结果

测验卷			因变量
简单对比(a)			数学平均分
级别 1 和级别 3	对比估算值		−22.500
	假设值		0
	差分(估计−假设)		−22.500
	标准误		2.965
	Sig.		.000
	差分的 95%置信区间	下限	−29.755
		上限	−15.245
级别 2 和级别 3	对比估算值		10.750
	假设值		0
	差分(估计−假设)		10.750
	标准误		2.965
	Sig.		.011
	差分的 95%置信区间	下限	3.495
		上限	18.005

a　参考类别＝3。

表 4-33 给出了三次测验之间数学平均分的比较结果，由于级别 1(测验卷 B1)与级别 4(测验卷 B3)的相伴概率值为 0.000 < 0.05，表示测验卷 B1 与测验卷 B3 的数学平均分有显著性差异；同样方法，我们可以看出测验卷 B2 与测验卷 B3 之间的数学平均分也有显著性差异。

表 4-34　不同测验卷之间数学平均分的检验结果

源	平方和	*df*	均方	*F*	Sig.
对比	2303.167	2	1151.583	65.493	.000
误差	105.500	6	17.583		

表 4　34 表明，从方差分析的角度看，由于相伴概率值 0.000 <

0.05，因此，不同测验之间的数学平均分的差异显著。

表 4 35　边际均值估算(1)：不同学校数学平均分的边际均值

学校	均值	标准误	95%置信区间	
			下限	上限
A1	111.500	2.421	105.576	117.424
A2	104.833	2.421	98.909	110.757
A3	102.000	2.421	96.076	107.924
A4	101.000	2.421	95.076	106.924

表 4-36　边际均值估算(2)：不同测验卷数学平均分的边际均值

学校	均值	标准误	95%置信区间	
			下限	上限
B1	86.250	2.097	81.120	91.380
B2	119.500	2.097	114.370	124.630
B3	108.750	2.097	103.620	113.880

表 4-37　不同学校之间数学平均分的两两比较

(I)学校	(J)学校	均值差值 (I−J)	标准误	Sig.	95%置信区间	
					上限	下限
A1	A2	6.6667	3.42377	.099	−1.7110	15.0443
	A3	9.5000(*)	3.42377	.032	1.1223	17.8777
	A4	10.5000(*)	3.42377	.022	2.1223	18.8777
A2	A1	−6.6667	3.42377	.099	−15.0443	1.7110
	A3	2.8333	3.42377	.440	−5.5443	11.2110
	A4	3.8333	3.42377	.306	−4.5443	12.2110
A3	A1	−9.5000(*)	3.42377	.032	−17.8777	1.1223

（续表）

(I)学校	(J)学校	均值差值 (I J)	标准误	Sig.	95%置信区间	
					上限	下限
A3	A2	−2.8333	3.42377	.440	−11.2110	5.5443
	A4	1.0000	3.42377	.780	7.3777	9.3777
A4	A1	10.5000(*)	3.42377	.022	18.8777	−2.1223
	A2	−3.8333	3.42377	.306	−12.2110	4.5443
	A3	−1.0000	3.42377	.780	−9.3777	7.3777

基于观测到的均值。
* 均值差值在.05级别上较显著。

表4-37显示，四所学校中，学校A1的学生水平与学校A3、A4之间有着比较显著的差异，其他学校之间的学生数学水平没有显著性差异。

表4-38 不同测验卷之间数学平均分的两两比较

(I)测验卷	(J)测验卷	均值差值 (I−J)	标准误	Sig.	95%置信区间	
					下限	上限
B1	B2	−33.2500(*)	2.96507	.000	−40.5053	−25.9947
	B3	−22.5000(*)	2.96507	.000	−29.7553	−15.2447
B2	B1	33.2500(*)	2.96507	.000	25.9947	40.5053
	B3	10.7500(*)	2.96507	.011	3.4947	18.0053
B3	B1	22.5000(*)	2.96507	.000	15.2447	29.7553
	B2	−10.7500(*)	2.96507	.011	−18.0053	−3.4947

基于观测到的均值。
* 均值差值在.05级别上较显著。

表4 38显示，三套测验卷的平均分有显著差异，说明这三套高考数学模拟测验试卷并不等值。

第五章 试卷分析报告的基本模式

撰写试卷分析报告的目的是为了全面认识测验的效果，正确理解测验成绩传达的信息，指导教师更有效地开展教学，帮助学生更好地进入下一阶段的学习。根据撰写人的不同，试卷分析报告可以分为学生测验后反思、教师测验后分析与专业人员测验后系统分析等三种形式；根据测验性质与规模的不同，试卷分析报告又可以分为单元测验分析、学期测验分析与大规模考试（联考与统考等）分析。无论是哪种试卷分析报告，都需要从定性与定量两个角度展开，并综合运用各种统计量、统计图表对测验结果进行统计分析，这既是评价测验质量、测验成绩的基本方法，也是形成试卷分析报告的基本形式。

第一节　试卷分析报告的基本框架

根据试卷结构与设计目标，试卷分析应从整卷、题组、题目等三个不同层面有目的、有侧重地报告有关测验信息，进行有关统计推断，旨在较为全面、深入地展示测验的整体面貌。下面以某市某年中考数学质量分析数据为例进行说明。

一、整卷层面

测验卷整体设计与测验效果的宏观分析是试卷分析报告的核

心内容，它向教师与学生展现测验目标达成状况、测验结果的分布状态，有利于教师掌握面上的教学情况，也有利于学生了解自己在班级、年级、地区所处的位置。

整卷层面的分析通常包括试卷的考查内容、考查结构与考查方式的研究，测验成绩的一般数量指标（集中量数、差异量数等）、质量指标（信度、难度、区分度等）的介绍与分析，测验成绩分布状态的呈现与分析等。

1. 整卷考查目标、考查内容与题型结构分析

某市某年中考数学学科考试时间120分钟，卷面满分为150分。全卷共三大题，25小题，其中选择题10小题，填空题6小题，解答题9题。客观性试题共30分，占全卷总分的20%。试卷的具体结构与上年对比如表5-1。

表5-1统计显示，除了试题难度比例做了较多的调整，其中容易题增加了12个百分点，难题增加了6个百分点，相应地中等题减少了18个百分点，其他的目标结构、内容结构、题型结构等部分仅仅具有微小的波动。各项特征构成分布的纵向差异，经过χ^2检验，均不具有显著性（$p>0.05$），说明本年度的中考命题与上年度相比，较好地贯彻了“稳中有变”的指导方针。

2. 整卷的数量特征

整卷的数量特征包括两个部分，一部分是报告全体考生的整卷量化指标及其相关分析，另一部分是报告不同区域（如，行政区域、学校、班级等，根据实际需要选择适当的数据整理方法）的整卷量化指标及其相关分析。如果是大型考试（如中考、高考等），就需要向学校公布不同区域、不同学校的整卷分数的量化指标并给出相关分析。

下面仅仅给出第一部分的数据及其分析。

表5-2中，整卷的各个基本量化指标的计算都是根据全体考生的测验成绩由计算机完成。整卷难度值为0.60，说明全卷难度适中，符合中考学业水平考试的要求。

表 5-1 试卷各项考查指标结构与分布

结构		本年			上年		差异比较	
		题号	分值	百分比	分值	百分比	比例差	χ^2 检验
目标结构	了解	2,11	6	0.04	6	0.04	0.00	$p > 0.05$
	理解	1，3，4，5，8，12，13，17	30	0.20	36	0.24	−0.04	
	掌握	6，7，14，15，18，19，20，21，22，23，24(1)，25(1)	87	0.58	87	0.58	0.00	
	灵活运用	9，10，16，24(2～3)，25(2)	27	0.18	21	0.14	0.04	
内容大致结构	数与代数	1，4，5，6，10，11，13，17，19，21，22，25(2)	72	0.48	62	0.41	0.07	$p > 0.05$
	空间图形	2，3，8，9，12，14，15，20，23，24，25(1)	63	0.42	68	0.45	−0.03	
	统计概率	7，16，18	15	0.10	20	0.13	−0.03	
题型结构	选择题	1，2，3，4，5，6，7，8，9，10	30	0.20	30	0.20	0.00	$p > 0.05$
	填空题	11，12，13，14，15，16	18	0.12	18	0.12	0.00	
	解答题	17，18，19，20，21，22，23，24，25	102	0.68	102	0.68	0.00	
难度结构	容易题	1，2，4，5，6，7，8，11，12，14，15，17，18	51	0.34	33	0.22	0.12	$p > 0.05$
	中等题	3，9，10，13，19，20，21，22，23，24(1)，25(1)	78	0.52	105	0.70	−0.18	
	难题	16，24(2～3)，25(2)	21	0.14	12	0.08	0.06	

（说明：内容结构分类部分，有些题目综合了不同领域的数学知识，归类时，则以考查的主要知识点来划分）

表 5－2 整卷常规量化指标分析

分 类	考生人数	平均分	难度	标准差	区分度	信度
含往届生	114523	90.06	0.6004	36.51	0.60	0.92
去掉往届生	112911	89.96	0.5997	36.59	0.60	0.92

整卷的区分度达到 0.60，属于优秀级别，反映出该试卷对考生的能力进行了很好的区分，说明该份中考试卷很好地兼顾了高中招生的需求。

整卷的信度值为 0.92，说明该次中考数学测验卷所测内容的同质性程度很高，达到最好的标准化考试的水平。

3. 整卷的分数分布

表 5－3 统计显示，学生的成绩分布较为理想，其中成绩在 0～50 分的学生占总人数的 18.8%，约有 57.19% 的学生成绩在 90～139 分，试卷较好地区分出不同能力层次学生的数学学习水平，反映出该地区初中生数学学习情况。

表 5－3 全体考生数学成绩按分数段的频数统计

分数段	人数	百分比(%)	累积百分比(%)
0	54	0.05	0.05
1～9	1105	0.96	1.01
10～19	3298	2.88	3.89
20～29	4977	4.35	8.24
30～49	6077	5.31	13.55
40～49	6009	5.25	18.80
50～59	6024	5.26	24.06
60～69	6140	5.36	29.42
70～79	6819	5.95	35.37
80～89	7848	6.85	42.22

(续表)

分数段	人数	百分比(%)	累积百分比(%)
90～99	8593	7.5	49.72
100～109	9935	8.68	58.40
110～119	14309	12.49	70.89
120～129	21315	18.61	89.50
130～139	11352	9.91	99.41
140～149	658	0.57	99.98
150	10	0.01	99.99
合　计	114523	99.99	

根据表 5－3，利用 EXCEL 软件制作考生成绩分布直方图如图 5－1。

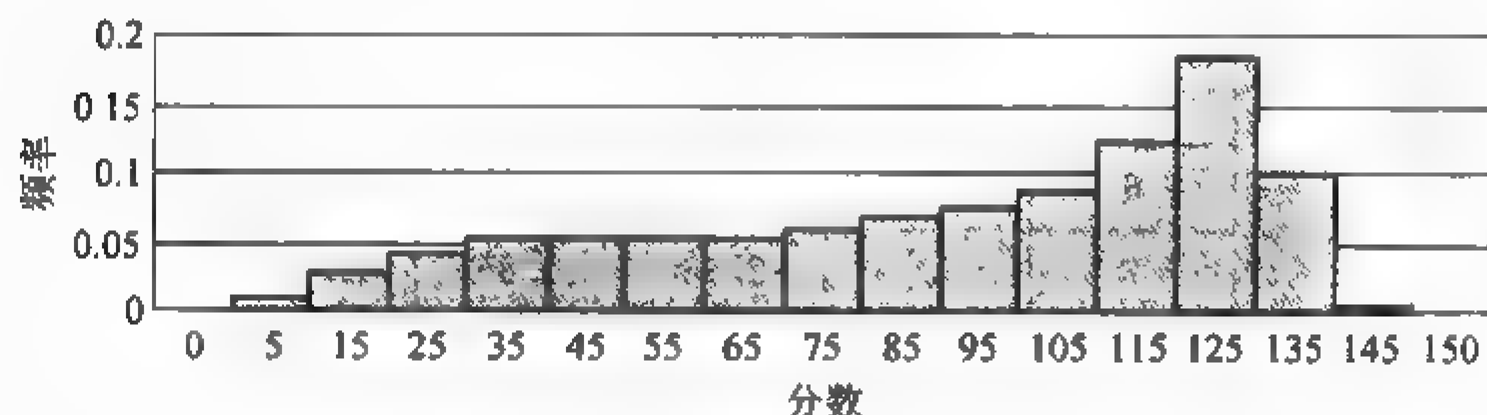


图 5－1　考生数学成绩分布直方图

图 5－1 显示，考生的考分频率分布直方图呈负偏态，说明数学考试的全卷难度适中略偏易。考生成绩没有出现双峰现象，说明数学学习两极分化现象没有凸显。

二、题组层面

1. 题组的整体数量指标

表 5　4 统计显示，选择题整体最简单，其次是填空题，解答题的难度最大。总体而言，三种题型的区分度都达到优秀标准。

表5 4　分题组的量化指标

题　型	满　分	平均分	难　度	标准差	区分度
选择题	30	24.17	0.81	6.20	0.41
填空题	18	13.40	0.74	4.05	0.41
解答题	102	52.50	0.51	31.47	0.64
合　计	150	90.07	0.60	36.51	0.60

表5-5　各个试题的量化指标

题　型	试题号	满分	平均分	难度	标准差	区分度
选择题	1	3	2.78	0.93	0.78	0.22
	2	3	2.69	0.9	0.92	0.27
	3	3	1.98	0.66	1.42	0.45
	4	3	[illegible].51	0.84	1.11	0.49
	5	3	2.7	0.9	0.9	0.34
	6	3	2.32	0.77	1.25	0.58
	7	3	2.56	0.85	1.06	0.35
	8	3	2.19	0.73	1.33	0.58
	9	3	1.85	0.62	1.46	0.48
	10	3	2.59	0.86	1.04	0.35
填空题	11	3	2.54	0.85	1.08	0.51
	12	3	2.77	0.92	0.79	0.20
	13	3	2.07	0.69	1.35	0.74
	14	3	2.51	0.84	1.11	0.41
	15	3	2.41	0.80	1.20	0.45
	16	3	1.1	0.37	1.45	0.16

（续表）

题　型	试题号	满分	平均分	难度	标准差	区分度
解答题	17	9	7.19	0.80	3.31	0.67
	18	9	6.97	0.77	2.91	0.52
	19	10	5.27	0.53	4.27	0.91
	20	10	6.9	0.69	3.69	0.78
	21	12	7.98	0.67	4.31	0.79
	22	12	6.31	0.53	5.51	0.95
	23	12	6.29	0.52	4.48	0.85
	24	14	4.09	0.29	3.58	0.54
	25	14	1.5	0.11	2.06	0.26

根据表 5－5 的统计信息，结合不同统计指标可以进行深入的分析。

2. 试题编排情况

根据表 5－5 中试题的难度状况，利用 EXCEL 软件制作整卷试题难度编排动态曲线图，如图 5－2。从整体上看，整卷试题的编排顺序呈现出由易到难、逐步递进的结构。其中，第 1～10 题为选择题，第 11～16 题为填空题，第 17～25 题为解答题，开卷中的第 1 题、

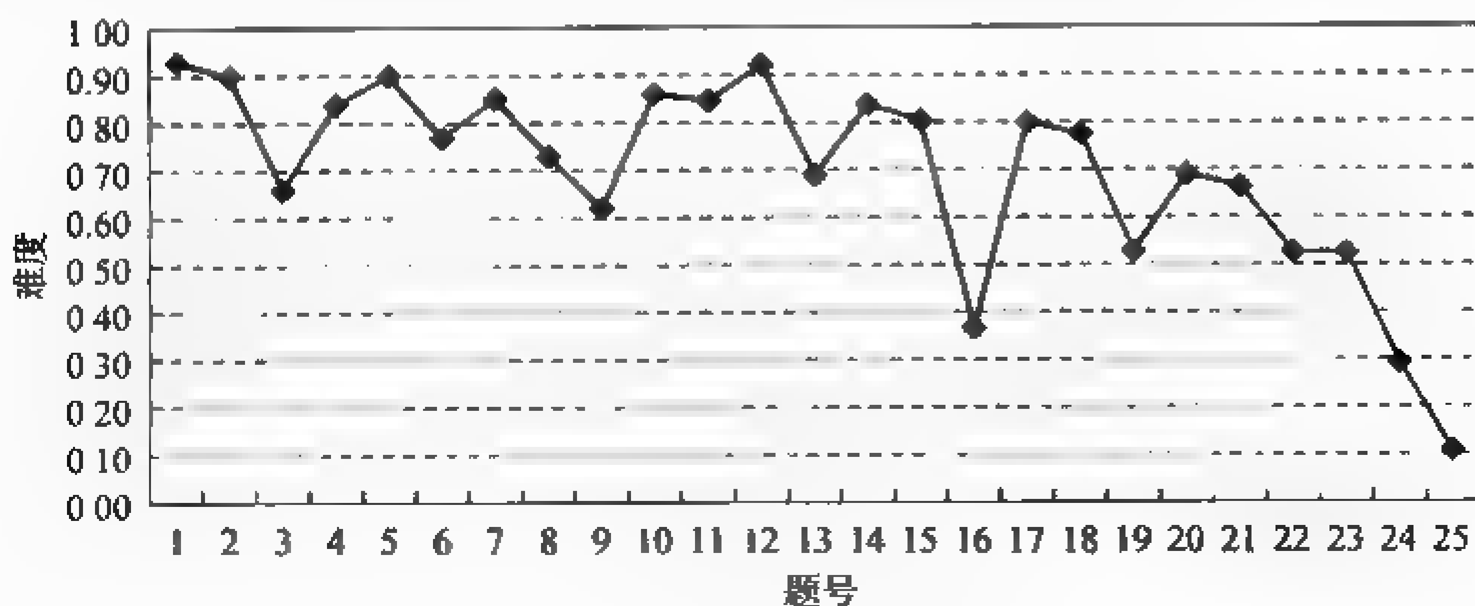


图 5－2　整卷试题难度编排动态曲线

题型转换的起始题(第11、16题)相对较为容易,每种题型中都有相对较难的试题;整卷中最难的两道题放置在最后(第24、25题),难度值分别达到0.29与0.11。整体而言,整卷的试题编排合理,有利于考生以良好的心理状态答卷并发挥自己的最好水平。而且全卷的容易题、中等题、难题的分值比例约为51∶78∶21,也部分说明全卷难度总体适中。

稍嫌不足的是,整卷的选择题部分波动平稳,基本上都属于容易题,难度梯度不明显,成为考生得分的主要部分。

3. 试题区分度状况

根据表5-5中试题的区分度状况,利用EXCEL软件制作整卷试题区分度情况动态曲线图,如图5-3。从整体上看,试卷中偏难试题(如第16、24、25题等)与偏易试题(如第1、2、12题等)的区分度效果一般,而中等难度的试题(如第19、22、23题等)的区分度效果非常理想。整卷的区分度高达0.60,说明整卷的区分度指标很优秀。试卷中,区分度小于0.2的试题只有第16题,相应的区分度值是0.16,具体情况在试题分析部分应给予详细分析。

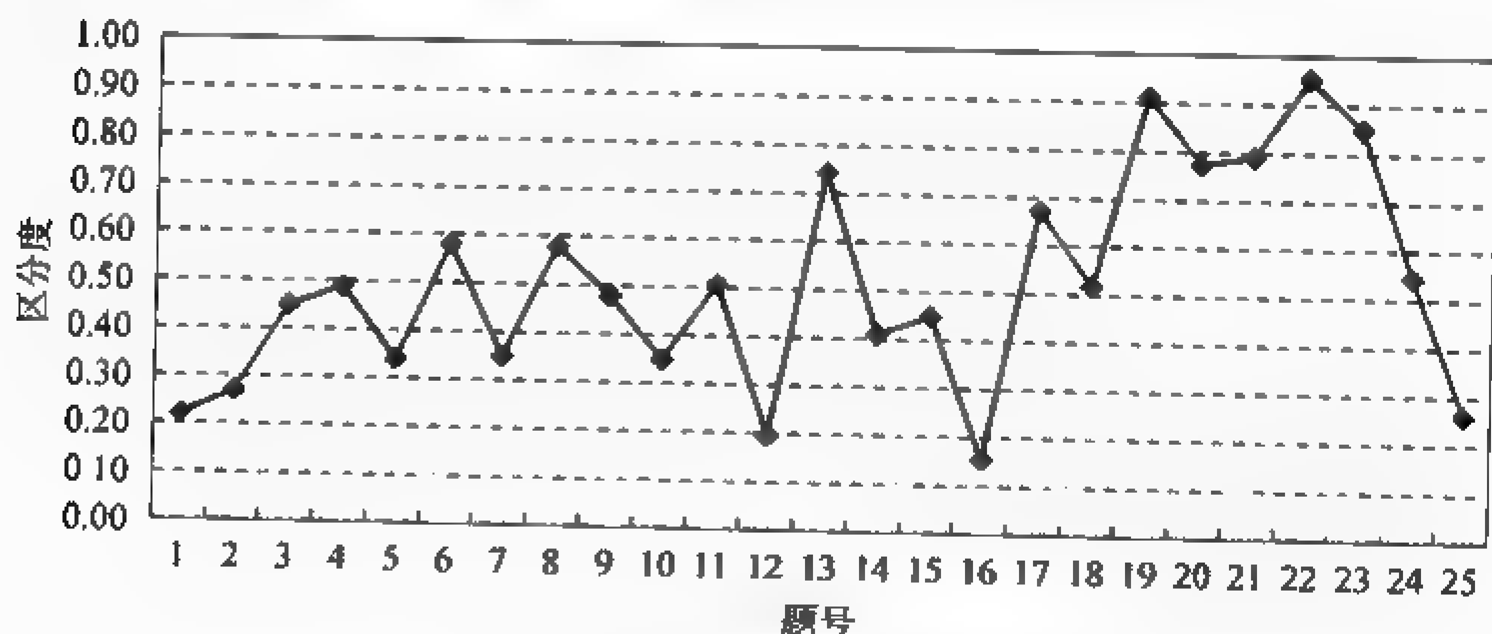


图5-3 整卷试题区分度情况动态曲线

三、试题层面

由于各个试题的具体数量指标在题组部分已经给出,因此试题层面

主要进行定性分析,从测验目标、测验内容、试题的命制方式、考生的答题情况等多个角度进行深层次细致的讨论,并作出相应的事实判断。

1. 试题命制情况

试题的质量问题既可能来源于试题命制方面,也可能来源于考生群体水平及其发挥状况。在分析试题质量时,从命制的角度进行深入分析,既有助于对教学内容的理解与把握,也有助于对考查方式的使用与指导,还有助于掌握命题技术。

对每个试题命制情况的分析,可以从考查知识点、试题来源、试题设计的题型使用情况(是否发挥了题型的优势)、题目的功能(对教师教的考查、对学生学习状况的考查、实际考察效果)等方面逐题展开研讨,具体格式可以参照表5-6进行,也可以不用列表,分试题逐一进行阐述,详略情况由分析者自行把握。最后应给出以后命题的改进建议。

表5-6 试题来源、题型使用及其功能

题号	知识点	试题来源	题型使用	题目的功能	
				考查学生的学习	考查教师的教学
1					
2					
3					
…	…	…		…	…
24					
25					

【例5-1】 试根据表5-6的模式分析某次高一测验卷中的一道试题:"如图5-4,在正方体$ABCD-A_1B_1C_1D_1$中,E、F为棱AD、AB的中点。(1)求证:EF∥平面CB_1D_1;(2)求证:平面$CAA_1C_1\perp$平面CB_1D_1。"

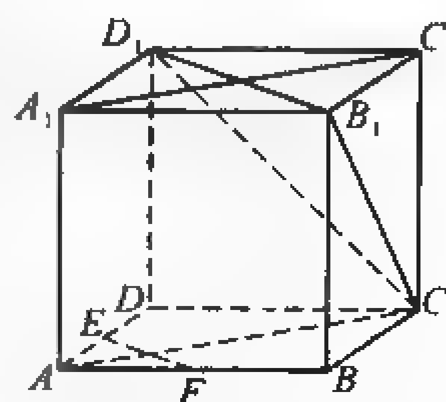


图5-4

具体分析如下。

1. 试题来源与主要知识点

本试题根据北师大版高中必修2教材第49页A组习题7改编，考查的知识点主要有：空间中线面、面面的位置关系。

2. 题目的功能

(1) 考查教师教的方面，重点是引导教师重视审题的教学，在几何教学中重视通过实际模型，认识点、线、面间基本的位置关系，强调观察角度的选取和对图形的分解，逐步加强对推理和思维严谨性的要求。

(2) 考查学生学的方面，要求学生掌握以正方体为模型的简单的线面平行与垂直关系的性质与判定的方法与技能。

3. 题型使用

本题要求考生书面写出详细解答过程，能够较为全面地检查出考生在必修1立体几何部分核心知识的掌握情况，题型应用合理。

2. 考生答题情况

测验核心目的之一是检查考生的学习掌握情况，因此，测验结束后对考生答题情况展开深入分析成为试卷分析的重要组成部分。

对每道试题考生答题情况的分析包括考生的整体答题状况（对平均分、难度、区分度、标准差等指标的具体分析）、典型解答方法、最优解答方法、典型错误、对教学的启示等方面的内容，具体可以参照表5－7进行深入讨论。

表5－7 每道试题考生答题情况分析基本模式

题号与题目					
基本量化指标	满分：	平均分：	标准差：	难度：	区分度：
典型解法					
最优解法					
典型错误及错因分析					
对教学的启示					

第二节　三种常见测验的试卷分析报告基本模式

大多数一线教师习惯于上试卷讲评课，但不习惯动笔撰写试卷分析报告，因为教师很熟悉试卷中试题内容与考生的答题情况，而且上试卷讲评课约束相对较少，可以即兴发挥；但撰写试卷分析报告的工作量大，尤其需要用到教师相对陌生的教育测量学、教育统计学与教育评价学的知识，教师感到为难。

其实，在教师经常遇到的单元测验、学期测验中，试卷分析的基本方法很简单，也很容易掌握，一旦形成撰写试卷分析报告的习惯，不仅花费时间不多，而且对提高教学的针对性和学生自我认识水平等都有很大帮助。

下面主要介绍单元测验、学期测验与联考或统考等三种常见测验的试卷分析报告的基本模式，具体运用时，教师可以根据具体情况适当删减。

一、单元测验

单元测验的试卷提倡由教师自行命制，或者由教师根据现成的试卷结合自己所教学生的实际学习状况修改后形成。

单元测验的功能主要定位于检测某单元知识的学与教的状况，关注的是某位教师与其任教班级在某个单元知识方面的教学现状，相应地，测验后的试卷分析就应关注该单元教学具体细节的处理与掌握情况，并由教师和学生共同完成分析报告。

1. 教师撰写部分

教师撰写的试卷分析报告主要包括学习内容与考查内容、班级整体测验情况分析、题组整体测验情况分析、试题测验情况分析、小结等五部分内容。

学习内容与考查内容的报告着重突出单元学习知识、技能和思想方法的重点，突出测验是为教学服务，学什么就考什么，并检查测验内容的覆盖率，确保基于该单元测验的分析具有一定的可靠性与有效性。

由于单元测验属于标准参照考试，区分度不需要过度强调；而且样本较小，测验信度与结构效度的计算结果仅作参考，因此，基于班级的单元测验可以不报告有关测验卷的信度、结构效度与试题区分度等量化指标。

在向全班同学报告测验成绩时，为了去除给学生排队的嫌疑，同时又方便学生了解自己的测验成绩在班级整体中的位置，建议教师给出班级测验成绩的茎叶图。由于班级人数较少，教师手绘茎叶图很简单，另外利用 SPSS 软件制作茎叶图也非常简便，具体请阅读第三章第二节茎叶图部分。

在分析题组情况时，既可以详细汇报每道题的常规量化指标，也可以整体汇报题组的常规量化指标，视测验情况而定。然后简要分析每个题组的考查目标与实际达成情况。

试题分析部分不需要面面俱到，大部分主要集中在两端：最优解答方法与暴露问题最多的试题剖析。通过分析发现产生错误的原因，才能真正以考促学。学生产生错误是概念法则不清，还是计算能力薄弱；是单项知识没掌握，还是综合运用所学知识解决实际问题的能力不强，仔细分析其出错原因，有助于教师及时调控教学策略和方法。另外，低年级可能需要在规范审题、解题、作图等方面进行引导。

最后应简要小结测验的效果与下一单元的教学基础，以及需要注意的问题。

具体操作可以参照表 5－8 的模式进行，这里以人教版九年级数学第 22 章一元二次方程单元测验为示例简要说明。

表 5-8　九年级上学期数学一元二次方程单元测验分析报告

撰写人(教师):________ 测验时间:________ 撰写时间:________

第一部分:学习内容与考查内容

<table>
<tr><th colspan="2"></th><th>学习内容</th><th>考查内容与承载的试题</th><th>分数比例</th></tr>
<tr><td rowspan="6">本单元内容</td><td rowspan="3">知识点</td><td>1. 对一元二次方程的认识</td><td></td><td></td></tr>
<tr><td>2. 一元二次方程的解法:
配方法、公式法、因式分解法</td><td></td><td></td></tr>
<tr><td>3. 实际问题与一元二次方程</td><td></td><td></td></tr>
<tr><td rowspan="2">主要技能</td><td>1. 配方的基本步骤与技巧</td><td></td><td></td></tr>
<tr><td>2. 求根公式的使用特点与要求</td><td></td><td></td></tr>
<tr><td>思想与方法</td><td>降次,化归</td><td></td><td></td></tr>
</table>

第二部分:班级整体成绩

<table>
<tr><td>班级:</td><td>班级总人数:</td><td>实考人数:</td><td>缺考人数:</td></tr>
<tr><td>平均分:</td><td>中位数:</td><td>难度:</td><td>标准差:</td></tr>
<tr><td>全班测验成绩茎叶图</td><td colspan="3"></td></tr>
</table>

(续表)

第三部分:题组答题分析

	第一大题 (如:选择题)	第二大题 (如:填空题)	第三大题 (如:解答题)
平均分			
难度			
标准差			
简要评析			

第四部分:试题答题分析

最优(最有创意)的试题解答及其分析	
典型错误、错误最多的试题及其错因分析	

第五部分:小结

对本单元教学的评价	关于教法
	关于指导学生
对下单元教学的启示	关于教法
	关于指导学生

2. 学生撰写部分

测验的核心目的之一是帮助学生认识自己某单元知识的学习状况,使学生能诊断出学习中的薄弱环节,以便及时查漏补缺,明确下一阶段的努力方向,因此,每次单元测验后组织学生撰写单元测验分析是一种很好的学习方法。

学生撰写单元测验分析时,可以从本单元知识掌握情况、错因

分析、下阶段学习目标等几个方面展开。

本单元知识掌握情况分析帮助学生认清自己学习的整体状态，哪些内容掌握得好，哪些内容掌握得不好，客观地评价自我。

如果测验没有得到满分，说明该单元学习有一些问题需要引起注意。错因分析部分引导学生进行自我诊断，具体反思答题错误是由智力因素引起还是非智力因素引起，通过自我评价来端正学习态度、改进学习方法、提高学习效率，达到自我教育的效果。

分析完出错原因后，最好让学生衡量自己的实际水平与测验分数之间的差距，明确自己是超水平发挥或是发挥失常，通过这种诊断，让学生看到自己可能达到的目标，制定学习的近期与远期目标，养成良好的自我规划的习惯。

当学生通过对试卷的系统分析、总结、反思后，他们意识到了自己存在的问题和不足。此时，教师应引导他们根据自身情况订出提高成绩的措施和方案，为下一阶段的学习拟定目标和策略。

为了引导学生更有效地完成测验后的反思，具体操作可以参照表 5－9 的模式进行，这里还是以人教版九年级数学第 22 章一元二次方程单元测验为示例简要说明。

表 5－9　九年级上学期数学一元二次方程单元测验后的反思报告

撰写人（学生）：__________ 测验时间：__________ 撰写时间：__________

第一部分：本单元知识掌握情况

知识点	承载试题题号	分值小计	实得分数	得分率
1．一元二次方程的概念				
2．用配方法解一元二次方程				
3．用公式法解一元二次方程				
4．用因式分解法解一元二次方程				
5．根的判别式				
6．根与系数的关系				

（续表）

7. 运用一元二次方程解应用题

第二部分：错因分析

出错试题题号	主要错误原因(每题限选至多3项)										
	完全不会	知识遗忘	审题出错	概念错误	推理出错	计算出错	公式用错	格式出错	表达不规范	来不及做	其他
	改正：										
	自编类似试题与解答：										
	改正：										
	自编类似试题与解答：										

第三部分：小结(知识掌握情况，学习方法使用情况，疑难与困惑、下阶段学习目标、具体措施等)

对本单元学习的评价	值得肯定部分	1
		2
		3
	学习薄弱部分	1
		2
		3
下单元学习目标	知识掌握方面	
	学习方法方面	
	具体措施	

在实际教学实践中，如果能够长期引导学生填写表5-9，还可

以发现学生在学习态度、学习方法、学习习惯等方面存在的问题，更有利于教师针对性地开展教学。

二、学期测验

学期测验往往指的是一个学期结束时的测试，通常由学校安排统一考试时间，全年级学生在规定的时间内完成相应科目的测验；测验结束后，由全年级的教师分科统一阅卷；阅卷结束后，汇总所有数据，并统一进行试卷分析；在此基础上，各班科任教师形成所任班级的试卷分析报告，并把测验情况反馈给学生。

学期测验的功能主要包括两个方面：一方面是检测某学期整体学与教的状况，关注的是年级整体的学与教的水平；另一方面，也比较各个班级的教学情况，以便进行教学评价。相应地，学期测验后的试卷分析除了关注该学期学科教学的处理与掌握情况外，还需要进行班级之间的教学情况对比分析。因此，学期测验分析报告应由年级学科备课组的教师共同完成。

学期测验试卷分析报告主要包括考查内容分布、年级整体测验情况分析、各个班级测验成绩分析、题组整体测验情况分析、试题测验情况分析、小结等六部分内容。

考查内容分布分析着重突出一个学期所学各章内容的考查结构、知识点分布，知识、技能和思想方法的考查重点，并检查测验内容的覆盖率，确保基于该学期测验的分析具有较好的可靠性与有效性。

虽然学期测验属于标准参照考试，区分度不需要过度强调，但由于样本较大（一个年级的学生往往有数百人），检验试题的区分度还是必要的；另外，学期测验对测验信度与结构效度的要求也相应高于单元测验，因此，也应提供相应的测验卷的信度与效度指标。由于年级总人数较多，建议用 EXCEL 或 SPSS 软件制作给出学期测验成绩的频数分布直方图，以方便学生了解自己在年级中所处的位置。

在统计各个班级的测验成绩时，除了公布本次测验的平均分外，还应公布入学分班时每个班的平均成绩。为了规范体现班级之

间的差异，应运用单因素方差分析法分别检验每次测验后班级平均分之间的差异是否显著（具体方法请参考第四章例4-11）。由于班级之间的平均分往往有波动，波动是否在正常范围内，应使用χ^2检验进行显著性分析。给出各个统计指标的数值后，应简要分析统计量的实际意义，以方便师生阅读与理解。

在分析题组情况时，只需要基于全年级的整体水平详细汇报每道题与每个题组的常规量化指标，而基于每个班级的相应每道题、每个题组的常规量化指标由各个科任教师完成，不体现在全年级学期测验分析报告中。

试题分析部分应逐题展开，这样为每个班的试题分析提供参照。

最后应简要小结学期测验的预测目标与实际效果的吻合度，下个学期的教学基础与需要注意的问题。

具体操作可以参照表5-10的模式进行，这里以人教版九年级上学期数学期末测验为示例简要说明。

表5-10　九年级上学期数学期末测验分析报告

撰写人（备课组）：________ 测验时间：________ 撰写时间：________

第一部分：考查内容

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="2">教学时数分配</td><td colspan="2">考查内容分配</td><td colspan="2">差异比较</td></tr>
<tr><td>课时数</td><td>课时比例</td><td>分值</td><td>百分比</td><td>比例差</td><td>χ^2检验</td></tr>
<tr><td rowspan="6">本学期内容</td><td>第21章　二次根式</td><td>9</td><td>15%</td><td></td><td></td><td></td><td rowspan="5"></td></tr>
<tr><td>第22章　一元二次方程</td><td>13</td><td>21%</td><td></td><td></td><td></td></tr>
<tr><td>第23章　旋转</td><td>8</td><td>13%</td><td></td><td></td><td></td></tr>
<tr><td>第24章　圆</td><td>17</td><td>28%</td><td></td><td></td><td></td></tr>
<tr><td>第25章　概率初步</td><td>14</td><td>23%</td><td></td><td></td><td></td></tr>
<tr><td>小计</td><td>61</td><td>100%</td><td></td><td></td><td></td><td></td></tr>
</table>

（续表）

第二部分：全年级成绩统计

年级总人数：	实考人数：	缺考人数：	
平均分：	最高分：	最低分：	标准差：
难度：	信度：	效度：	区分度：
全年级测验成绩频数分布直方图			

第三部分：各个班级测验成绩统计

班级	入学分班(或上次测验)情况			本次测验情况			差异比较		
	平均分	排名	平均分差异检验	平均分	排名	平均分差异检验	均分差	排名差	χ^2检验
1									
2									
3									
4									

第四部分：题组答题分析

		平均分	难度	标准差	区分度
第一大题	第1题				
	第2题				
	……				
	第8题				
	小　计				
第二大题	第9题				
	第10题				
	……				
	第14题				
	小　计				

（续表）

第三大题	第 15 题				
	第 16 题				
	……				
	第 20 题				
	小　计				
合　　计					
简要评析					

第五部分：试题答题情况分析

题号	分析内容	
1	考查要点	
	最优解答	
	典型错误	
	对教学的启示	
2	考查要点	
	最优解答	
	典型错误	
	对教学的启示	
……	……	……
20	考查要点	
	最优解答	
	典型错误	
	对教学的启示	

第六部分：小结

本学期全年级教学情况总体评价	优点：
	不足：
下个学期全年级教学目标规划	目标：
	措施：

除了年级备课组给出学期测验的试卷总体分析报告外，各个班级的科任教师还需要结合任教班级的考生答题情况进行针对性分析，在此基础上组织全班学生进行试卷分析，上好测验讲评课。

学期测验中，学生同样需要撰写相应的测验分析或反思。学生撰写学期测验分析时，可以借助单元测验的分析模式进行，这里不再赘述。

三、联考或统考

联考即联合考试，它存在多种组织形式，常见的主要有两种：教学水平相当的校与校之间民间性质的联考，区与区之间的半官方性质联考。统考即统一考试，它具有官方性质，往往由有关教育行政部门组织进行。此处主要探讨具有官方性质的联考与统考的试卷分析报告的撰写特征。

联考与统考的命题工作都由专业人员完成，这类考试既具有标准参照测验的性质，也兼顾常模参照测验的特征；施测后的分析既关注本校学生的答题情况，又涉及跨校或跨区域的成绩比较研究。因此，撰写联考与统考的试卷分析报告时，与前述单元测验、学期测验的侧重点有较大不同。

联考与统考的试卷分析报告应包括试卷质量分析、试题质量分析、考生总体测验成绩统计与答题情况分析和不同区域（学校）测验成绩对比分析等内容。

试卷质量分析包括测验目的定位准确性分析、考查内容有效性分析、测验结果可靠性分析、试卷编排合理性分析、评分标准制定与实施的确切性分析等部分，每部分的量化指标计算、定性分析方法具体可以参考第二章相关内容。

试题质量分析包括试题编写的科学性与规范性分析、试题的难易度分析、试题的鉴别力分析、答题的典型错例分析、对教学的指导作用等部分，具体可以参照第二章相关内容进行研究。

考生总体测验成绩统计与答题情况分析包括测验成绩分布状

况、测验分数的组织与概括、基于测验分数进行的统计推断、考生卷面答题情况的质性分析等方面，具体可以参照第三、四章的相关内容，也可以参照单元测验、学期测验的相关部分分析。

不同区域（学校）测验成绩对比分析应关注多个侧面、多个层次，力求客观、全方位地反映不同区域（学校）的教学面貌，为下阶段的教学与管理提供坚实依据。

下面以某市某五个区域联考数学测验成绩为例，简要介绍对不同区域测验成绩进行比较的常见方法。

表 5-11　某五区联考数学测验成绩统计

区域	平均分	难度	标准差	区分度	前 1/3 平均分	后 1/3 平均分	及格率
A	79.02	0.53	34.59	0.58	119.69	32.82	42.13
B	83.2	0.55	34.81	0.58	124.32	35.59	47.25
C	89.36	0.60	40.03	0.65	135.01	37.32	59.18
D	94.62	0.63	34.49	0.56	134.85	44.43	62.41
E	101.4	0.68	35.07	0.55	139.82	50.46	71.1

从表 5-11 中，不仅可以了解每个区域的整体状况，还可以了解按照考生分数平均分成三段的前后两段考生的平均分和及格率状况，根据表 5-11 还可以制作平均分分布折线图，如图 5-5，以更形

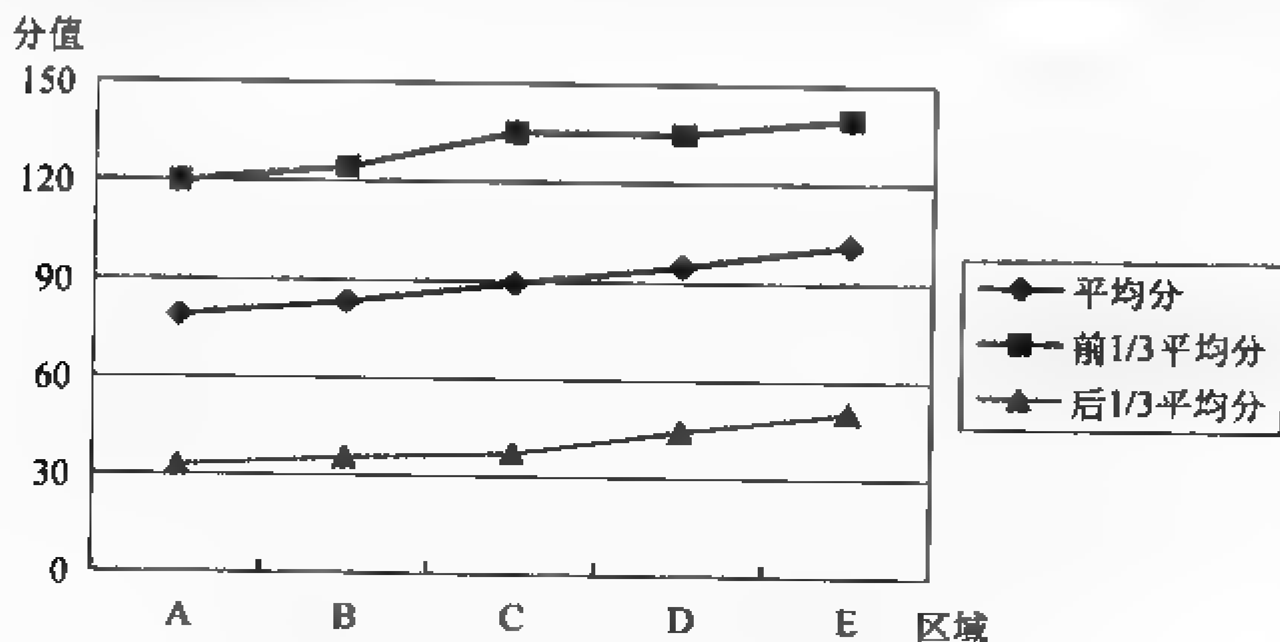


图 5-5　某五区联考数学测验三类平均成绩折线统计图

象地反映各个区域的测验成绩。其中区域C的标准差、区分度等两个指标值都最大，虽然前1/3平均分高于区域D，但由于后1/3平均分低于区域D很多，因此总平均分明显低于区域D，这说明区域C学校间、学生间两极分化现象明显，需要给予高度关注。

表5 12 某五区联考数学测验高分段(前10%)测验成绩统计

区域	学校总数	学生总数	学校分布		学生分布	
			学校数	占所属区总数比例	学生数	占所属区总数比例
A	25	9201	10	40.00%	20	0.22%
B	11	2551	1	9.09%	1	0.04%
C	51	8904	14	27.50%	127	1.43%
D	33	8371	16	48.48%	123	1.47%
E	37	12607	31	83.78%	456	3.62%
小计	157	41634	72	45.86%	727	1.75%

表5-12统计显示，虽然区域A的三项平均分指标都明显偏低，但优等生比例强于区域B，而且全区40%的学校都有成绩拔尖的学生，如何发挥优等生的榜样作用，需要深入研究。区域B的优生培养比较薄弱，需要重点突破。区域C的优生非常集中，虽然优生总人数不少，但是分布的学校较为集中，体现出学校之间的明显差异。区域E优生人数多且分散，全区83.78%的学校都有优生领跑，再加上学困生人数比例相对较小，因此，整体优势非常明显。

表5-13统计显示，区域C教学质量薄弱学校所占比例最大，在32所学校中比例高达59.38%！而同时教学质量良好的学校所占比例相对而言差异不算太大。其次，区域A教学质量薄弱学校所占比例不容乐观，需要具体问题具体分析。总体而言，区域E的教学质量良好的学校多，而教学质量薄弱的学校几乎没有，整体优势明显。

另外，还可以用单因素方差分析法具体比较区域内学校与学校

表5 13 平均成绩排名前20%、后20%学校在五区分布统计

区域	学校总数	前20%				后20%			
		前10%学校数	百分比(%)	前20%学校数	百分比(%)	后10%学校数	百分比(%)	后20%学校数	百分比(%)
A	25	0	0	2	6.25	2	12.50	8	25.00
B	11	0	0	0	0	2	12.50	3	9.38
C	51	4	25.00	9	28.13	10	62.50	19	59.38
D	33	3	18.75	8	25.00	2	12.50	2	6.25
E	37	9	56.25	13	40.63	0	0	0	0
小计	157	16	100	32	100	16	100	32	100

间平均分的差异、区域与区域间平均分的差异，比较方法参见第四章第三节，这里不再赘述。

附录 1 概化理论简介

虽然经典测量理论在心理与教育测量中得到广泛应用，其作用和地位越来越显著，但是，经典测量理论的先天不足也极大地困扰着使用者，该理论的局限性日益突出。一些测验研究者从深入分析测验误差的来源和结构出发，应用方差分量分析方法来辅助测验研究，在经典真分数理论的基础上创建了从宏观上研究测验性质的新理论——概化理论。

第一节 概化理论对经典真分数理论的拓展

经典真分数理论把实测分数简单地划分为真分数和误差分数（$X = T + E$），并且把误差分数看成是随机误差，对不同来源的误差分数不做进一步的分解与探讨。概化理论认为测量误差中既有系统误差，也有随机误差，并从“测什么”与“怎么测”的角度具体界定测量目标、影响和制约测量目标的各种因素、测量分数、影响和制约测量分数的各种因素，然后运用方差分析法同时讨论各种因素可能产生的误差对测量结果的影响。

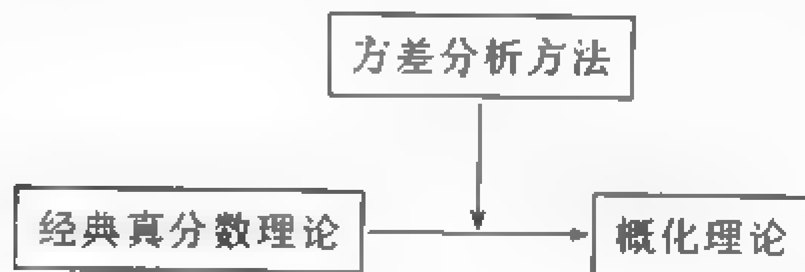
经典真分数理论一般假设考生总体的能力状况服从正态分布，进行测验时强调根据随机抽样理论选取考生样本，强调样本选取的代表性。概化理论也认为测验要观察的考生一般均抽样来自某一

个总体，但对考生总体的分布没有明确的规定。

经典真分数理论假定平行测验测量的是同一种能力，因此所有平行测验的平均分、方差与协方差等均相等。概化理论则将一份测验试卷看作是一个由无数道试题组成的全域（区别于测量对象的总体）的一个样本，只要两份测验卷是从同一个试题全域中随机抽得的，所构成的测验就是平行测验。

经典真分数理论是在施测后分析数据，确定误差值的大小，并进行相关的分析。概化理论则提出“先概括、后决策”的两步工作方法，先在一定的测量条件下设计并进行试验性的测试，按照试测所得的数据估计各种来源的方差分量，然后再根据相应方差分量决定的指标去考查：当改变测量条件的某些方面时会出现何种结果，据此加以判断，做出今后应如何去控制、改进测量精度的优化决策。

概化理论与经典真分数理论的关系可以用下面的框图简单概括。



第二节　概化理论的基本概念

概化理论在突破与拓展经典真分数理论时形成了一系列的概念、原理与方法，构成该理论的基本体系。以下基本概念是理解概化理论的基础，下面结合教育测验加以介绍。

1. 测量侧面、侧面水平与观察全域

教育测验的根本目的是推测考生具备的知识与技能、方法与能力的程度，因此，测验研究的对象就是考生。然而，考生在测验中获得分数的高低，除了决定于其自身对知识技能方法的掌握情况、个

人能力的高低外，还会受到许多其他因素影响，如试题的难度是否与考生的能力匹配、测验时间的规定是否恰当、评分标准的制定是否合理、评分教师对评分标准的把握是否一致，等等。在开展研究时，如果把考生组的某种能力看成测量目标，那么，影响和制约考生发挥真实水平的各种因素就称为测量侧面，每个影响因素称为一个侧面，如，试题侧面、测验时间侧面、评分标准侧面、评分教师侧面，等等。

一个侧面可以有不同的水平。例如，在试题侧面中，如果一份测验卷由 n 道试题构成，那么这 n 道试题就代表了试题侧面的 n 个水平；在评分教师侧面中，如果有两名教师参加阅卷，那么这两名教师就代表评分教师的 2 个水平。

测量侧面还有随机侧面与固定侧面之分。随机侧面是指该侧面的水平数是无限的，在测量分析中所使用的水平是从该侧面所有水平中随机抽取的一个样本。例如，在数学测验中，可以将试题侧面看成是随机侧面，一次测验所使用的测验卷就构成该随机侧面的一个样本，测验卷中的试题数就是该随机侧面的水平数。固定侧面是指该侧面的水平数是一个确定值（如 a），在测量分析中所使用的水平就是这 a 个水平。例如，在数学测验中，可以将测验内容侧面看成是固定侧面，如果一次测验只涉及"一次函数的概念、图象与性质、简单应用、与方程不等式结合的综合应用"等知识，那么该固定侧面就分为这 4 个水平。

由于测验总是在一定条件下进行的，因而总是存在一定数量的测量侧面，测量中可能存在的测量侧面的全体就构成测量侧面的总体，称为观察全域。

2. 概化研究

概化理论认为各个测量侧面是测量误差的来源。例如，考生 A 在一次数学测验中获得 78 分，那么影响考生 A 得分的主要侧面有：考生 A 在相应考查内容上的实际数学能力，试题设计，评分教师对评分标准的把握等，可能还有更多的测量侧面需要研究。为了确定

不同测量侧面对考生分数的影响效应，需要确定不同侧面效应的方差，这种不同侧面方差就成为方差分量。

概化研究就是根据测量目的与用途，先确定测量目标与测量侧面的结构，再设计收集资料的方案，根据设计方案进行试验性测试，以调查各个侧面的采样对考生测验分数的影响，并提供尽可能多的测量误差来源的信息，即确定各个侧面方差分量的大小。

概化研究把观测分数的总体方差分解成测量目标方差、不同测量侧面方差、测量目标与测量侧面交互作用方差、不同测量侧面交互作用方差、交互作用与其他不明变异来源混杂效应的残差方差等部分，其一项主要任务就是用方差分析等方法来估出各方差分量的估计值，这种估计值应反映出观察全域中可能存在且又可以实际观察的各种影响因素（侧面）所造成的方差。

3. **拓广全域**

概化理论通过对观察全域中各个测量侧面进行概化研究，来发现观察全域中可能存在的问题，从而提出对观察全域的有关侧面的修改方案，以形成一个新的全域，这个新的全域就是拓广全域（也称作概化全域），它包括研究者希望把研究结果推广而至的所有侧面数及其每个侧面的相应水平数，一般是观察全域的子集。

4. **决策研究**

决策研究就是利用概化研究提供的信息，通过增加侧面的水平数、将侧面固定等策略，寻找减少误差、提高测量精度的良好的测验设计，使得某个或某些侧面对测验的影响或误差最小。这个研究过程称为决策。

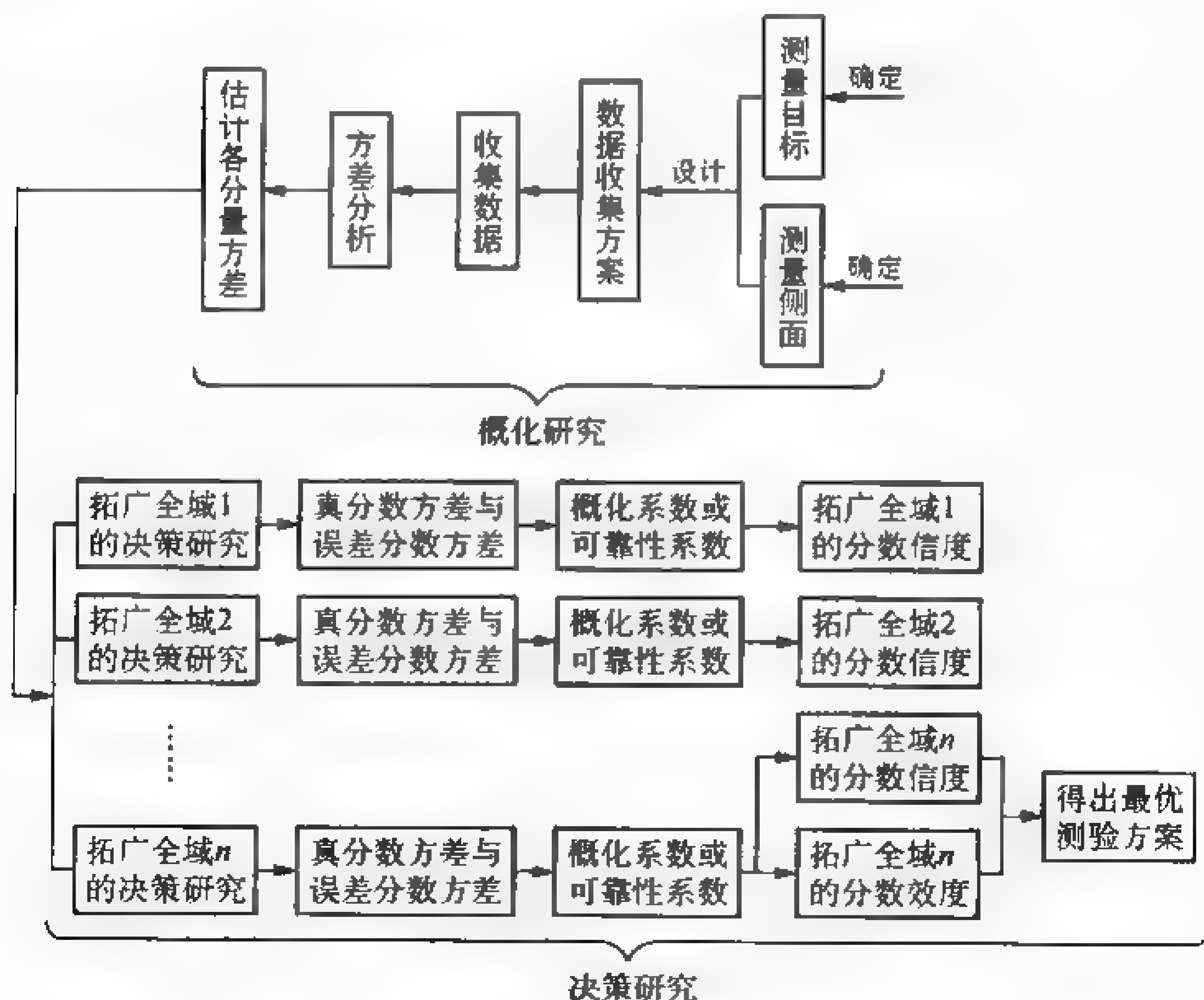
根据研究者对测验结果作出处理的方式来划分，决策分为相对决策和绝对决策两种。相对决策是指把某一考生的分数与其他考生进行比较而做出决策，例如，常模参照测验的结果解释。绝对决策是指把考生的答题情况与教育、教学的客观标准进行比较而做出决策，例如，标准参照测验的结果解释。

第三节　概化理论的基本原理

1. 概化理论分析流程图

运用概化理论研究测验时，研究过程分为依次进行的概化研究、决策研究两步。概化研究就是定量估计观察全域中测量目标、测量侧面等形成的各种测量误差方差，为决策研究提供分析数据。决策研究就是利用概化研究所获结论，去考察如何提高测量精度，从而作出优化决策，实现预控调节。

概化理论研究分析过程的流程图如下。



2. 概化研究阶段收集测验数据方案的设计

数据收集方案的设计类型包括交叉设计、嵌套设计和交叉与嵌

套混合设计三大类型。

交叉设计是指，根据测量目的确定出所有的测量目标与测量侧面，每个测量目标在每个测量侧面的所有水平上均被测量，所有类型的测量数据都被收集。教育测验中常见的交叉设计主要有单侧面设计、两侧面设计两种。例如，某市某年中考的单侧面交叉设计要求所有考生应该完成所有试题，数据收集的方式可以用 $s\times i$ 表示，其中 s 代表考生数，i 代表试题数，见表1；如果该市该年中考采用双侧面交叉设计，则两侧面一般是试题和评分教师，要求所有考生均完成所有试题，同时，所有评分教师对所有试题进行评分，数据收集的方式可以用 $s\times i\times j$ 表示，其中 s 代表考生数，i 代表试题数，j 代表评分教师数，见表2。

表1　单侧面交叉设计 $s\times i$

收集的实测分数		试题侧面			
		试题1	试题2	……	试题 i
被试总体	考生1	x_{11}	x_{12}	……	x_{1i}
	考生2	x_{21}	x_{22}	……	x_{2i}
	⋮	⋮	⋮	⋮	⋮
	考生 s	x_{s1}	x_{s2}	……	x_{si}

表2　双侧面交叉设计 $s\times i\times j$

收集的实测分数		评分教师1				…	评分教师 j			
		试题1	试题2	……	试题 i	…	试题1	试题2	……	试题 i
被试总体	考生1	x_{111}	x_{121}	……	x_{1i1}	…	x_{11j}	x_{12j}	……	x_{1ij}
	考生2	x_{211}	x_{221}	……	x_{2i1}	…	x_{21j}	x_{22j}	……	x_{2ij}
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	考生 s	x_{s11}	x_{s21}	……	x_{si1}	…	x_{s1j}	x_{s2j}	……	x_{sij}

嵌套设计是指，根据测量目的确定出所有的测量目标与测量侧面，然后把某个侧面的各个水平分别包含到另一个侧面的各个水平

之中，再针对性地收集测量数据的方法。教育测验中常见的嵌套设计主要有单侧面嵌套设计、两侧面嵌套设计两种。例如，某市某年中考英语口语的单侧面嵌套设计要求部分考生完成一部分试题，另一部分考生完成另一部分试题，数据收集的方式可以用 $i:s$ 表示，其中 i 代表试题的分类数，s 代表每部分试题考生数，当 $i=2$ 且 $s=3$ 时的示例见表 3；如果该市该年中考英语口语采用双侧面嵌套设计，同样，两侧面一般是试题和评分教师，要求所有考生均完成所有试题，同时，部分评分教师对部分试题进行评分，数据收集的方式可以用 $s\times(i:j)$ 表示，其中 s 代表考生数，i 代表试题的部分数，j 代表评分教师数，当 $s=6$，$i=2$，$j=2$ 时示例见表 4。

表 3　单侧面嵌套设计 $i:s$

收集的实测分数		第一部分试题		第二部分试题	
		试题 1	试题 2	试题 3	试题 4
被试总体	考生 1	x_{11}	x_{12}		
	考生 2	x_{21}	x_{22}		
	考生 3	x_{31}	x_{32}		
	考生 4			x_{43}	x_{44}
	考生 5			x_{53}	x_{54}
	考生 6			x_{63}	x_{64}

表 4　双侧面嵌套设计 $s\times(i:j)$

收集的实测分数		评分教师 1		评分教师 2	
		试题 1	试题 2	试题 3	试题 4
被试总体	考生 1	x_{111}	x_{121}		
	考生 2	x_{211}	x_{221}		
	考生 3	x_{311}	x_{321}		
	考生 4			x_{432}	x_{442}
	考生 5			x_{532}	x_{542}
	考生 6			x_{632}	x_{642}

交叉与嵌套混合设计是指把交叉设计、嵌套设计综合使用的一种方法，它一般用于有三个测量侧面及其以上的情况。

三种设计类型中，交叉设计的数据信息是最丰富的，而纯嵌套设计的数据信息是最简单的。理论上讲，设计的测量侧面越多，每个侧面涉及的水平数越多，那么对测验的分析就越完善；但是，对于后续的统计分析而言，困难就会越大，甚至可能无法进行。

收集完数据进行方差分析时，可以借助计算机统计软件，如GLM、GENOVA、SAS 中的 VARCOMP 等进行，数据处理较为简便。

3. 决策研究阶段的研究方法简介

决策研究阶段，首先要根据概化研究提供的各种来源误差方差的估计值，在原设计方案收集的数据范围内，对各个测量侧面做出不同的调整，得到一些拓广全域，然后形成一些新的测验方案。调整方法主要有三种。

方法一是将一个或几个随机侧面改为固定侧面（至少保留一个随机侧面）。例如，研究评分教师评分的信度时，可以将试题看成是固定侧面，将评分教师看成随机侧面，研究改变评分教师数和评分教师构成等对测验结果的影响。

方法二是调整一个或几个测量侧面的水平数。例如，调整试题数，或调整评分教师数等。一般地，增加水平数意味着增加测量的重复数，可以达到提高测量精度的目的。

方法三是改变测量数据的收集方法，主要是把交叉设计的数据部分或全部地改为嵌套设计，达到减少投入、简化测量的目的。

对于变化了的各种新测验方案，决策研究给出了两个比较优劣的误差指标：相对误差方差与绝对误差方差。在误差指标的基础上，又进一步给出了测验精度的两个综合指标：概化系数（用于衡量常模参照测验质量）与可靠性系数（用于衡量标准参照测验质量）。可以根据这些指标判定测验方案的质量。

第四节　运用概化理论应注意的问题

就目前发展状况来看，运用概化理论分析测验行为时需要注意以下问题。

1. 考生样本的选取应具代表性，以确保数据分析的可靠性

应用概化理论全面分析测验性质时，首先需要设计测验，在此基础上采集测验数据，分析各种测验误差方差，再在此基础上分析比较各种可能的测验方案，从而优选出最佳的新测验方案。因此，概化理论的分析基础是测验数据，为了保证概化分析结果的可靠性，必须满足样本数据的代表性，即充分保证考生样本的代表性。

2. 施测条件的控制应注意前后一致性

任何测量都依赖于特定的施测条件，施测条件中的测量目标、测量侧面、测量侧面水平数等的变化都会引起测验误差来源、测验误差大小、测验信度等的变化，从而造成测验分数解释范围的变化，因此，运用概化理论分析测验行为与测验结果时，应尽量保持施测条件的前后一致性。

3. 测量侧面的确定应兼顾测验组织实施、数据统计技术等的可操作性

概化理论的重心是分析各种测量误差的来源，并尽量减小测量误差。从理论上讲，测量侧面和测量水平数越多，对测验的分析就越完善。然而，测验侧面过多，不仅会有施测组织和实施的困难，后期数据统计、数据分析的工作量也会过大，甚至可能导致无法完成数据分析工作。因此，在进行测验设计时，应充分考虑测验的可操作性，以确保测验实施与分析的正常进行。

附录2 试题反应理论简介

虽然概化理论对经典真分数理论做出许多改进与发展，但是概化理论所使用的测量模型与经典真分数理论并没有质的不同，它所探讨的还是停留在外部条件对考生作答的种种影响，研究的是如何控制外部条件造成的误差来源问题，仍然没有把考生作答情况与试题测量性能、考生实际能力有机地结合起来。试题反应理论则克服了经典真分数理论、概化理论的这一不足，另辟新路，成为现代测量理论中最具代表性的一种。

第一节 试题反应模型的3个基本假设

试题反应理论建立在潜在特质理论基础之上。考试、测验总是要考查测量人的某种内部心理特性，如智力、能力等，由于所要考查的对象都不是直接可以观察到的，因此被统称为潜在特质。试题反应理论研究的是，这种潜在特质水平与试题特征如何联合起来共同决定考生在该试题上的答对概率（简称答对率），至于潜在特质的心理结构或特征如何，并不探讨。

为了从数学的角度刻画试题特征、潜在特质水平与试题答对率之间的关系，试题反应模型构建出试题特征曲线及其相关函数模型，这些模型的构建通常需要满足以下3个基本假设。①

① 丁向英主编，教育测量与统计，郑州大学出版社2004年1月，30页。

1. 潜在特质的单维性

即考生的某一测验结果只取决于一种潜在特质或能力，其他能力的影响可忽略不计。

2. 试题解答的局部独立性

这个假设包括两个方面，一是考生解答某道试题时，不受其他试题的影响，即某一考生整份测验卷的答对率等于每道试题答对率之积。二是考生与考生之间在进行试题解答时是相互独立、互不影响的。

3. 数学模型的恰当性

即选取的试题反应模型应与测验数据之间的拟合良好。在进行试题分析之前，必须对数学模型的拟合度进行统计检验。

第二节　试题特征曲线与试题特征函数

一般而言，一道编制质量良好的试题应体现出这样的特征：考生的测验总分越高，考生在该试题的答对率也越大。为了更好地揭示考生能力与试题特征如何共同确定试题的答对率，美国著名测量学家洛德用考生的测验总分作为考生能力水平 θ 值，以 θ 为自变量，考查每道试题在每个 θ 值上考生的答对率，并描点连线得出试题特征曲线，发现试题特征曲线是一条中心对称的 S 形曲线，然后用逻辑斯递(Logistic)函数来描述该曲线。

逻辑斯递函数的表达式如下：

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta - b_i)}}。 \qquad (1)$$

其中，$P_i(\theta)$是能力水平为 θ 的考生在试题 i 上的答对率，c_i 是试题 i 的猜测度(即猜测答对的概率)，a_i 是试题 i 的区分度，b_i 是试题 i 的难度。

一般地，当试题 i 确定后，该试题的参数 a_i、b_i、c_i 就变成定值，函数 $P_i(\theta)$就随能力水平 θ 的变化而变化。当试题不同时，试题参

数也不同，函数式(1)也不同。函数式(1)对应的曲线形状如图 1。

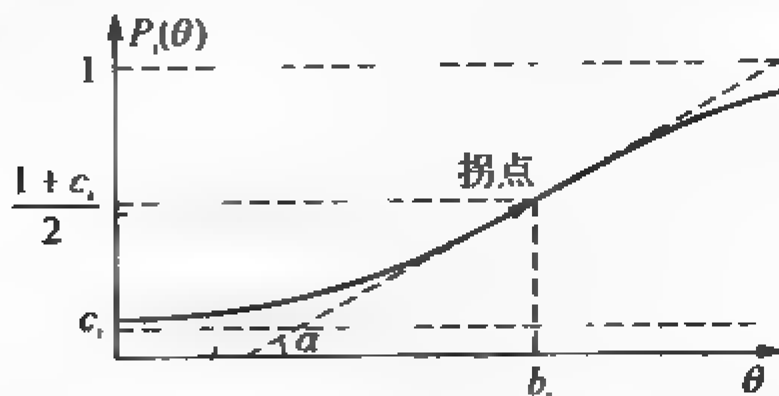


图 1　试题 i 的特征曲线

在图 1 中，自变量 θ 表示的是考生的某种潜在特质，习惯上采用标准 Z 分数表示，一般取值范围为$(-3, 3)$。$P_i(\theta)$随 θ 值的增大而增大，其图象位于两条平行线 $P_i(\theta)=c_i$，$P_i(\theta)=1$ 之间，其中 $P_i(\theta)=c_i$ 表示即使是能力水平极其低下的考生，试题答对率也有 c_i，因此称 c_i 为试题的猜测度，显然 c_i 越小越好；$P_i(\theta)=1$ 表示当考生的能力水平越来越高时，试题答对率越来越接近 1。试题特征曲线的拐点是$\left(b_i, \frac{1+c_i}{2}\right)$，它也是曲线的中心对称点；当 b_i 值增大时，曲线向右平移，只有潜在特质 θ 高的考生才能在新试题上获得相同的答对率，因此，称 b_i 是试题 i 的难度。过拐点作试题特征曲线的切线，记切线的倾斜角为 α，则 $a_i=\sqrt{2\pi}\tan\alpha$，$\alpha$ 越大，则曲线越陡峭，说明 θ 在 b_i 值附近稍有变化，则答对率差异就很大，即该试题把 b_i 值附近的考生进行了精细区分，这就是称 a_i 是试题 i 的区分度的含义。

根据试题特征函数，人们既可以对试题的质量作出评价，也可以估计考生在某一试题上的能力水平。需要指出的是，试题特征函数中的三个参数 a_i、b_i、c_i 虽然沿用了经典真分数理论中的名称，但是定义的角度与方式有了质的变化，应注意区分。

在实际应用中，有些试题的猜测度很小，为了研究的方便，就令 $c_i=0$，这时试题特征函数中的参数就变成两个，函数式(1)就简化为

$$P_i(\theta)=\frac{1}{1+e^{-1.7a_i(\theta-b_i)}}。\tag{2}$$

式(2)称为试题的双参数模型。

还有些测验，不仅所有试题的猜测度很小，可以忽略不计，而且所有试题的区分度也彼此接近，这时可以令 $a_i=1$，式(2)可以进一步简化为如下形式：

$$P_i(\theta)=\frac{1}{1+e^{1.7(\theta-b_i)}}。\quad (3)$$

式(3)称为试题的单参数模型，也被称为拉希(Rasch)模型。

第三节　试题反应模型与模型中的参数估计

试题反应理论的基本思路是针对测验中各种类型的试题，构造出不同的试题特征函数，用这些函数来揭示考生的试题答对率与考生能力水平、试题特征间的定量关系。

如果已知测验卷的每道试题参数，那么根据考生的作答反应，运用试题反应理论可以估计考生的能力水平。具体方法如下：

【例 1】 假设某测验卷由 5 道客观题组成，每道试题都是两级计分(分对与错两种情况)，各道试题参数与考生甲的作答反应情况如表 1。试估计考生甲的能力水平。

表 1　考生甲的作答情况表

试题	试题参数			试题答对率	甲作答情况
	区分度	难度	猜测度		
1	a_1	b_1	c_1	$P_1(\theta)$	1
2	a_2	b_2	c_2	$P_2(\theta)$	1
3	a_3	b_3	c_3	$P_3(\theta)$	0
4	a_4	b_4	c_4	$P_4(\theta)$	1
5	a_5	b_5	c_5	$P_5(\theta)$	1

分析：本题的试题反应模型是单维三参数模型，根据局部独立性假设，出现作答情况 $u=(1,1,0,1,1)$ 的概率就等于这五道试题答对率的乘积，这是一个关于能力水平 θ 的函数，称为似然函数，它的具体表达式为

$$L(u\mid\theta)=P_1(\theta)P_2(\theta)[1-P_3(\theta)]P_4(\theta)P_5(\theta)。\qquad(4)$$

考生甲的能力水平估计值就是使得似然函数 $L(u|\theta)$ 值达到极大值时自变量 θ 的取值。利用似然函数达到极值时估计参数 θ 值的方法称为极大似然估计法。具体计算步骤如下：

首先对函数(4)两边取对数，得到对数似然函数如下：

$$\ln L(u\mid\theta)=\ln P_1(\theta)+\ln P_2(\theta)+\ln[1-P_3(\theta)]+\ln P_4(\theta)+\ln P_5(\theta)。\qquad(5)$$

其次，令式(5)的一阶导函数值为零，即可得到关于 θ 的非线性似然方程如下：

$$\frac{1.7a_1(P_1-c_1)}{P_1}+\frac{1.7a_2(P_2-c_2)}{P_2}-\frac{1.7a_3(P_3-c_3)}{1-P_3}+\frac{1.7a_4(P_4-c_4)}{P_4}+\frac{1.7a_5(P_5-c_5)}{P_5}=0。\qquad(6)$$

然后采用牛顿—拉夫逊迭代法求解关于 θ 的非线性似然方程(6)，这一步需要借助计算机来完成，即可得到考生甲的能力水平估计值 θ_0。

如果运用试题反应理论指导测验编制，那么试题参数的估计是必不可少的工作。参数估计的思路如下：

第一步，首先编制一份测验卷（共有 n 道试题，假设每道试题都是 0—1 两级计分），并组织 m 个考生参加测验，获取所有考生每道试题的测验分数，得到如表 2 所示的分数矩阵。

表 2　分数矩阵表

		试题及作答情况				总分
		1	2	…	n	
考生	1	1	1	…	0	x_1
	2	0	1	…	1	x_2
	…	…	…	…	…	…
	m	1	0	…	1	x_m

第二步，假设考生的能力水平已知，估计每道试题参数。这时，可以依次估计每道试题参数，基本思路如下：

根据表 2 中的列分数向量，类比例 1，得到试题 i 的似然函数 $L_i(u|a_i, b_i, c_i)$，将对数似然函数 $\ln L_i(u|a_i, b_i, c_i)$分别对参数 a_i、b_i、c_i 求偏导，并令其值为零，可以得到关于 a_i、b_i、c_i 的三元方程组，然后采用牛顿—拉夫逊迭代法求解关于 a_i、b_i、c_i 的非线性方程组，即可得到试题参数的估计值。

如果考生能力水平未知，则需要同时估计考生能力水平和每道试题参数，这样需要估计的参数共有 $m+3n$ 个。解决思路如下：

首先根据分数矩阵表得到每个考生完成测验的对数似然函数 $\ln L_{ij}(u_{ij}\quad \theta_j, a_i, b_i, c_i)$，然后将 $\ln L_{ij}(u_{ij} \mid \theta_j, a_i, b_i, c_i)$ 分别对 θ_j $(j=1, 2, \cdots, m)$ 与 a_i、b_i、$c_i$$(i=1, 2, \cdots, n)$ 求偏导，并令其值为零，可以得到关于 $\theta_j$$(j=1, 2, \cdots, m)$、$a_i$、$b_i$、$c_i$$(i=1, 2, \cdots, n)$ 的 $m+3n$ 元方程组，然后采用牛顿—拉夫逊迭代法，从设定一套参数初值开始，经过反复迭代获得一组解序列，可以证明这组解序列最终收敛于方程组的真解。

第四节　试题反应理论的优点

试题反应理论的整个理论框架和分析方法与经典真分数理论、

概化理论有着质的不同，这也决定了试题反应理论具有以下独特的优点。

1．试题参数的跨群体不变性

在试题反应理论中，试题参数的不变性指衡量试题质量的各个参数（难度、区分度与猜测度等）不依赖于参加测验的考生样本，即试题参数不因考生样本的不同而不同。对于试题 i，根据试题特征函数 $P_i(\theta)$，能力水平为 θ_0 的考生答对试题 i 的概率仅仅与考生能力水平相关，而与考生在哪个群体无关，也与考生所在群体的大小无关，还与考生在群体中所处的位置无关，即 $P_i(\theta)$ 的大小由 θ 值唯一确定。试题反应理论的这一优点为建设大型题库、编制各种测验提供了理论依据。

2．能力参数的跨测验不变性

在试题反应理论中，能力参数的不变性体现为在同一个能力量表上，考生能力的大小与所施测的特定测验卷无关，即对考生能力的估计不因测验改变而改变。在试题反应理论中，真分数 T 可以定义为 $T=\frac{1}{n}\sum_{i=1}^{n}P_i(\theta)$（又称为测验特性函数），其中 n 是测验卷中的试题总数，$P_i(\theta)$ 是试题 i 的答对率，θ 是能力参数。当测验卷中的试题参数确定后，考生真分数 T 就完全由能力参数 θ 来确定。无论测验卷的具体试题构成如何，只要满足试题参数不变，那么测验特性函数就是唯一确定的，即能力参数 θ 独立于测验所施测的具体试题样本，并决定考生在各个具体测验上的真分数值。试题反应理论的这一优点为针对不同水平的考生实施题目不同的测验、建设自适应测验奠定了理论和方法基础。

3．能力参数与试题难度参数的直接可比性

试题反应理论把试题难度参数定义为试题特征曲线上拐点的横坐标，即把难度看成能力尺度上的位置参数，这说明能力参数与试题难度参数位于同一度量系统上。同时，试题反应理论还直接用能力参数与难度参数的对比，即 $(\theta-b)$ 作为揭示考生答对试题概率

高低的根本原因，从而引导人们选择恰当难度的试题去最有效地开展测试，为测验等值、试题有偏性探查等问题提供了解决途径。

4. 试题区分度参数与难度参数的相互独立性

根据试题特征曲线，难度参数是曲线拐点的横坐标；而区分度参数由曲线在拐点处斜率决定，与拐点的位置无关，即与难度参数无关。试题反应理论的这一性质为在任何难度水平上选择高区分度试题提供了保证。

5. 试题与测验信息函数的引进

试题反应理论还引进了如下全新的概念：试题信息函数与测验信息函数。试题信息函数、测验信息函数分别定义如下：

$$I_i(\theta) = \frac{(P'_i(\theta))^2}{P_i(\theta)(1-P_i(\theta))}, \tag{7}$$

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta)。 \tag{8}$$

式(7)中，$I_i(\theta)$为试题 i 的信息函数，$P_i(\theta)$为试题 i 的答对率。式(8)中 $I(\theta)$表示整份测验卷的信息函数。

式(7)针对每个能力水平定量地刻画出试题难度、区分度、猜测度是如何共同决定试题的测试功能，它表明每道试题提供的信息量既与考生能力水平有关，也与试题自身性能特点相关，但与其他试题无关。因此，试题信息函数具有可加性，从而得到测验信息函数是所含全部试题信息函数之和，即式(8)。

相应地，试题测量标准误、测验卷测量标准误分别为

$$SE_i(\theta) = \frac{1}{\sqrt{I_i(\theta)}}, \tag{9}$$

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}}。 \tag{10}$$

从上述定义可以看出，试题反应理论中的测量标准误不仅与参测的试题性质有关，还与参测的考生能力水平有关。即用相同试题

对不同的考生施测，其测验误差并不相同。这就为准确估计每个考生能力水平提供了准确信息，也为控制不同能力水平考生的测量误差提供了标准，还为测验试题编制提供了一种新型的、切实可行的选题策略。

第五节　试题反应理论的局限

虽然试题反应理论得到广泛的重视，但是目前该理论的应用也存在许多困难，这主要源于该理论存在的一些局限性，主要体现在以下四个方面。

1. 单维性假设不一定能得到满足

目前，常见的试题反应模型都有单维性假设，即假设只有一种能力起决定性作用，而其他能力可以忽略不计。但在教育测验中，人们对学科知识的理解与掌握需要靠平时知识的积累与学习，也应具有多方面的能力。因此，试题反应理论的单维性假设在应用时不一定能够得到满足，事实上，不少学者在研究成果中给出了这方面的案例，例如，指出语文高考测验不能满足单维性假设。[①]

当某些科目不满足单维性假设时，可以设想把整个科目分解成若干个分测验，使每一个分测验能满足必须得到满足的假设，再应用试题反应理论进行试题分析。这样做，又可能引起另外一些问题，如各个分测验题目量的大小，分测验之间分数的合成等，这些问题有待进一步解决。

2. 对数学模型与实测数据的拟合度要求较高

试题反应理论的核心是构建数学模型分析试题性质、考生能力水平与试题作答情况的关系，数学模型的好坏直接影响研究质量。

---

① 张敏强，刘晓瑜，项目反应模型的应用问题研究，心理学报，1998年10月，436—441页。

在实际应用中，如何选择适当的数学模型，如何检验实测数据与模型的拟合度，都是试题反应理论中备受关注的重大问题。

3. 计算工作量太大，计算过程复杂

根据前面的分析，试题反应理论的理论框架虽然比经典真分数理论、概化理论更为合理，但是涉及的计算原理专业性强，不易为普通教育者理解；且计算过程复杂，计算量太大，在普通教育教学领域的应用与推广具有一定的局限性。

4. 对主观性试题的测量与评价有待开发

目前，试题反应理论在单位特质测量与双歧评分试题研究方面已经比较成熟，但在多维特质测量、多级评分试题测试等方面的研究与应用还很有限，还远远不能满足各方面测验发展的需要。

主要参考文献

[1] 孙道德主编,概率论与数理统计,北京:人民教育出版社,2009年2月第1版。

[2] 崔允漷,王少非,夏雪梅主编,基于标准的学生学业成就评价,上海:华东师范大学出版社,2008年9月第1版。

[3] 沈有乾著,教育统计学,福州:福建教育出版社,2007年12月重印版。

[4] 张敏强主编,教育与心理统计学,北京:人民教育出版社,2007年8月第1版。

[5] 雷新勇著,考试数据的统计分析和解释,上海:华东师范大学出版社,2007年7月第1版。

[6] 华夏素质研究所课题组编,2006年全国中考数学评价报告,上海:华东师范大学出版社,2007年2月第1版。

[7] 漆书青著,现代测量理论在考试中的应用,武汉:华中师范大学出版社,2006年12月第1版。

[8] 凌云著,考试统计学,武汉:华中师范大学出版社,2006年12月第2次印刷。

[9] 王景英主编,教育统计学(第2版),北京:高等教育出版社,2006年10月第2版。

[10] 雷新勇著,大规模教育考试:命题与评价,上海:华东师范大学出版社,2006年4月第1版。

[11] 胡中锋主编,教育测量与评价(第二版),广州:广东高等教育出版社,2006年3月。

[12] 戴海崎、张峰、陈雪枫主编,心理教育测量,广州:暨南大学出版社,2006年1月。

[13] 华夏素质研究所课题组编,2005 年全国中考数学评价报告,长春:东北师范大学出版社,2006 年 4 月第 1 版。
[14] 顾海根编著,学校心理测量学,南宁:广西教育出版社,2005 年 10 月。
[15] 沈钢编著,教育统计与 EXCEL,杭州:浙江大学出版社,2004 年 12 月第 1 版。
[16] 余民宁著,教育测验与评量——成就测验与教学评量(第二版),台北:心理出版社,2004 年 10 月。
[17] 肖筱南主编,新编概率论与数理统计,北京:北京大学出版社,2004 年 6 月第 1 版。
[18] 杨晓明主编,SPSS 在教育统计中的应用,北京:高等教育出版社,2004 年 5 月第 2 版。
[19] 于向英主编,教育测量与统计,河南:郑州大学出版社,2004 年 1 月第 1 版。
[20] 吴明隆编著,SPSS 统计应用实务——问卷分析与应用统计,北京:科学出版社,2003 年 10 月第 1 版。
[21] Allen, W. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.
[22] R. L. Thorndike, & E. P. Hagen (1977), *Measurement and Evaluation in Psychology and Education* (*Fourth Edition*), copyright, by John Wiley & Sons, Inc.
[23] 张顺清,如何进行试卷分析初探,考试研究,2008 年第 6 期。
[24] 廖云霞,区分度在考试试卷分析中的应用,华中师范大学 2008 届硕士学位论文。
[25] 李伟,高中数学试卷质量分析的研究,华中师范大学 2008 届教育硕士学位论文。
[26] 马伟开,如何进行中学数学试卷分析,中学数学教学,2008 年第 1 期。
[27] 王建立,面向对象的通用试卷分析系统的设计与实现,山东大学 2006 届硕士学位论文。
[28] 胡维芳,论项目反应理论,高等理科教育,2005 年第 3 期(总第 61 期)。
[29] 王新,浅谈试后卷面分析的重要性,长春理工大学学报(社会科学版),2004 年 9 月。
[30] 尹传存、魏春梅,试卷质量的统计分析,曲阜师范大学学报,2004 年 7 月第

30卷第3期。
[31] 李方满,关于试卷与试题的分析,肇庆学院学报,2003年10月第24卷第5期。
[32] 李善富、秦毅,试卷分析设计探析,教学与管理,2003年7月。
[33] 赵天玉、陈圣滔,考试成绩与试卷质量的分析研究,江汉石油学院学报(社科版)2003年6月第5卷增刊。
[34] 葛正洪、尹相桂,试卷质量的综合评价,华北科技学院学报,2002年3月第4卷第1期。
[35] 张敏强,20世纪教育测量学发展的回顾与现状评析,教育研究,1999年11期。
[36] 张敏强、刘晓瑜,项目反应模型的应用问题研究,心理学报,1998年10月,30卷4期。
[37] 俞晓琳,项目反应理论与经典测验理论之比较,南京师范大学学报(社会科学版),1998年4月17卷第3期。

教师基本功丛书·数学教师卷

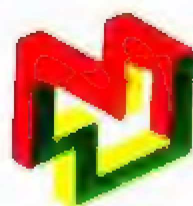
如何备好一堂数学课

如何上好一堂数学课

数学作业的设计与评价

数学学困生的转化

如何命数学题

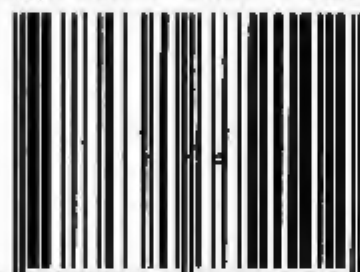


数学试卷分析方法

数学教育课题研究及论文撰写指导

多媒体数学课件制作

ISBN 978-7-5617-7210-2



定价：15.00元

www.ecnupress.com.cn